# FINITE DIFFERENCE METHODS FOR NON-LINEAR HYPERBOLIC SYSTEMS

John Ll. Morris

A Thesis Submitted for the Degree of PhD
at the
University of St Andrews

1968

# FINITE DIFFERENCE METHODS

# FOR

# NON-LINEAR HYPERBOLIC SYSTEMS

A thesis presented by

John Ll. Morris, B.Sc.,

to

the University of St. Andrews,

in application for

the degree of

Doctor of Philosophy

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

# DECLARATION

I declare that the following thesis is a record of research work carried out by me, that the thesis is my own composition, and that it has not previously been presented in application for a higher degree.

[redacted]

# PREFACE

In October 1962, I matriculated at the University of Leicester and read for a degree in Mathematics. In June 1965, I graduated with Second Class Honours in Mathematics. In October 1965, I was admitted as a full-time Research Student in the Department of Applied Mathematics of St. Salvator's College under the supervision of Dr. A.R. Mitchell.

# ACKNOWLEDGEMENTS

The Author is indebted to the Science Research Council for a Research Studentship held during the period of research and to Miss P.A. Cunningham for the expert typing of a major part of this thesis. The Author also expresses sincere appreciation for the many fruitful discussions with, and the constant encouragement by Dr. A.R. Gourlay.

# CERTIFICATE.

I certify that John Ll. Morris has spent nine terms in full-time research work under my direction, and is thus qualified to submitt the accompanying thesis in application for the degree of Doctor of Philosophy.

Research Supervisor.

# CONTENTS.

# CHAPTER I

# INTRODUCTION

### 1.1 Partial Differential Equations: Their existence and origin.

A partial differential equation is an equation which involves the rates of change of an unknown function with respect to variables whose dimensions are usually time or length. The term "order" of a partial differential equation will arise frequently and is defined to be the order of the highest derivative occuring in the partial differential equation. For example consider the partial differential equation

$$\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = 0 \tag{1.1.1}$$

where u is the unknown function and t, x and y are the variables having dimensions of time, length and length respectively. It is seen that the order of the highest derivative in (1.1.1) is two, and so (1.1.1) is said to be a second order partial differential equation. In the same manner, the equation

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x \partial y} - \frac{\partial^3 u}{\partial x^2 \partial y} - \frac{\partial^2 u}{\partial y^2} = 0 \tag{1.1.2}$$

is said to be of order three.

The order of a system of partial differential equations is defined similarly except that now the unknown quantity is a vector which may contain derivatives of unknown functions as dependent variables. For example

$$\frac{\partial \underset{\sim}{w}}{\partial t} + \frac{\partial \underset{\sim}{w}}{\partial x} + \frac{\partial \underset{\sim}{w}}{\partial y} = 0$$

where $\underset{\sim}{w} = \begin{pmatrix} u \\ uu_x \\ u^{\frac{1}{2}}u_{tt} \end{pmatrix}$ , say

is a first order system in $\underset{\sim}{w}$ .

In the following chapters we will deal in the main with first order systems of partial differential equations. In deriving numerical methods of solution we will always assume that we have been given such a first order system. Many scalar equations of higher order can of course be reduced to first order systems. For example the second order wave equation

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}$$

may be transformed to the first order system

$$\frac{\partial \underset{\sim}{w}}{\partial t} + A\,\frac{\partial \underset{\sim}{w}}{\partial x} = 0$$

by introducing new variables $w_1 = \frac{\partial u}{\partial t}$ and $w_2 = \frac{\partial u}{\partial x}$

and writing $\underset{\sim}{w} = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$ . A is the 2 x 2 matrix $\begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}$

This practice of reducing higher order scalar equations to systems of lower order is well known and will not be discussed further.

So far we have not stated how partial differential equations arise. The mathematical representation of the physical law, fundamental to a particular problem, often gives rise to a partial differential equation (or system of such equations) which thereafter is used to define the state of that problem. Because in most cases the physical law connects rates of change of some unknown function it is necessary to solve the partial differential equation in order to discover the behaviour of the unknown function. The solution of the differential equation in some region R is usually required to match certain initial and boundary conditions which are given as an essential part of the problem.

In the examination of partial differential systems the question naturally arises, "is there a solution?" and if there is one, "is it the only one?", that is, does a solution exist and if so, is it unique? In this thesis, the existence and uniqueness of the solution of the system of partial differential equations which we consider will always be assumed. In this context reference can be made to Lax [42], Conway and Smoller [8], Jeffrey and Tanuiti [36], Friedrichs [15], Bernstein [3], etc. In what follows we will consider non-linear and quasilinear differential equations as distinct from linear and semi-linear ones. It is advantageous therefore to distinguish between these classes by a series of examples and definitions.

Consider the non-linear partial differential equation

$$g(x, t, u, u_x, u_t, u_{xx}, u_{xt}, u_{tt}) = 0 \tag{1.1.3}$$

If (1.1.3) can be written as

$$a(x,t,u,u_x, u_t)u_{tt} + b(x,t,u,u_x, u_t)u_{xt} + c(x, t, u, u_x, u_t)u_{tt}$$
$$+ f(x, t, u, u_x, u_t) = 0$$

it is said to be *quasilinear* as it is linear with respect to the highest derivatives $u_{xx}$, $u_{xt}$, $u_{tt}$.

If (1.1.3) can be written as

$$A(x, t)u_{xx} + B(x, t)u_{xt} + C(x, t)u_{tt} + F(x, t, u, u_x, u_t) = 0$$

then it is said to be *semi-linear*, because the coefficients of $u_{xx}$, $u_{xt}$, $u_{tt}$ do not involve u or any of its derivatives. If together with this last form, F is a linear function of u and its derivatives, then (1.1.3) is said to be *linear*.

## 1.2 Classification of systems of partial differential equations.

We shall be concerned with a particular class of first order systems of partial differential equations namely the hyperbolic class and will not consider elliptic or parabolic classes. Accordingly the conditions for a first order system of partial differential equations to be hyperbolic will now be given.

Consider, first, in matrix notation, the first order system of partial differential equations

$$\frac{\partial u}{\partial t} + A\frac{\partial u}{\partial x} + B = 0 \tag{1.2.1}$$

where u is an m-dimensional column vector (the dependent variable) whose components are unknown functions $u_j$ ($j = 1, 2, \ldots, m$) of the two independent variables x and t. A is an m x m matrix whose elements $a_{ij}$ may be functions of x, t and u, and B is an m column vector with elements which may be functions of x, t and u. We require to derive conditions under which (1.2.1) is hyperbolic.

Consider now those curves known as <u>characteristics</u>. A characteristi of the system (1.2.1) is a curve, along which the value of u together with the equation (1.2.1) is not sufficient to determine the normal derivative of u to the curve. If we introduce the unit vectors $\mathbf{i}_s$ and $\mathbf{i}_\nu$ in the directions of the tangent and normal respectively at a point on a characteristic, where s is the arc length measured along the characteristic from a fixed origin to the point, then the derivative of a function u in the direction of the tangent to the characteristic is given by

$$(1.2.2) \qquad \frac{\partial u}{\partial s} = \underset{\sim}{i}_s \cdot \nabla u = \frac{\partial u}{\partial x}\frac{dx}{ds} + \frac{\partial u}{\partial t}\frac{dt}{ds}$$

where $\nabla u$ is the gradient of the function u. Similarly, the derivative of the function u along the normal is given by

$$(1.2.3) \qquad \frac{\partial u}{\partial \nu} = \underset{\sim}{i}_\nu \cdot \nabla u = -\frac{\partial u}{\partial x}\frac{dt}{ds} + \frac{\partial u}{\partial t}\frac{dx}{ds}$$

with suitable orientation of the (x, t) co-ordinate system. The position of the vectors may be seen more clearly with the aid of figure (1.2.1).

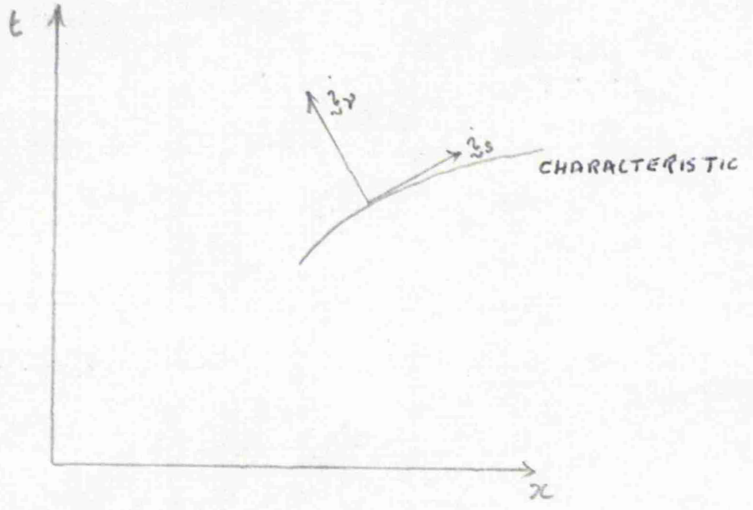


figure (1.2.1)

Solving (1.2.2) and (1.2.3) for $\frac{\partial u}{\partial x}$ and $\frac{\partial u}{\partial t}$ we obtain a unique solution provided

$$(1.2.4) \qquad \Delta = \begin{vmatrix} \frac{dx}{ds} & \frac{dt}{ds} \\ -\frac{dt}{ds} & \frac{dx}{ds} \end{vmatrix} = \left(\frac{dx}{ds}\right)^2 + \left(\frac{dt}{ds}\right)^2 \neq 0$$

But $\triangle = 1$ by virtue of $(ds)^2 = (dx)^2 + (dt)^2$. Hence we obtain

$$(1.2.5) \qquad \frac{\partial u}{\partial x} = \frac{\partial u}{\partial x}\frac{dx}{ds} - \frac{\partial u}{\partial \nu}\frac{dt}{ds},$$

and

$$(1.2.6) \qquad \frac{\partial u}{\partial t} = \frac{\partial u}{\partial s}\frac{dt}{ds} + \frac{\partial u}{\partial \nu}\frac{dx}{ds}.$$

By substituting for $\frac{\partial u}{\partial x}$ and $\frac{\partial u}{\partial t}$ from (1.2.5) and (1.2.6) into (1.2.1), we obtain

$$(1.2.7) \qquad (x_s I - t_s A)\frac{\partial u}{\partial \nu} + (t_s I + x_s A)\frac{\partial u}{\partial s} + B = 0,$$

where $x_s$, $t_s$ represent $\frac{dx}{ds}$ and $\frac{dt}{ds}$ respectively, and where I is the unit matrix. If the characteristic does not determine the normal derivative $\frac{\partial u}{\partial \nu}$ uniquely, the determinant of

$$(x_s I - t_s A)$$

must vanish, that is

$$(1.2.8) \qquad \left| x_s I - t_s A \right| = \left| A - \lambda I \right| = 0$$

where $\lambda = \frac{x_s}{t_s}$. We see therefore that every characteristic curve for the system (1.2.1) is a solution of the ordinary differential equation

$$\frac{dx}{dt} = \lambda$$

where $\lambda$ is one of the characteristic roots of the polynomial equation (1.2.8). By considering the left characteristic row vector $\underset{\sim}{v}$, say, corresponding to $\lambda$ and multiplying equation (1.2.7) on the left by this vector, (1.2.7) may be reduced to

$$(1.2.9) \qquad \underset{\sim}{v}(t_s I - x_s A) \frac{\partial u}{\partial s} + \underset{\sim}{v}B = 0,$$

which has the property that every derivative appearing in it is directed along the tangent to the characteristic.

If all the characteristic roots $(\lambda_1, \dots \lambda_m)$ of $A$ are real and have associated with them a full set of $m$ linearly independent characteristic vectors $(\underset{\sim}{v}_1, \dots \underset{\sim}{v}_m)$ then the system of partial differential equations (1.2.1) can be transformed into a canonical system

$$(1.2.10) \qquad \underset{\sim}{v}_k\left(\frac{\partial t_k}{\partial s} I + \frac{\partial x_k}{\partial s} A\right) \frac{\partial u}{\partial s} + \underset{\sim}{v}_k B = 0 \qquad k = 1, \dots m$$

of $m$ distinct equations involving differentiation in just one characteristic direction $\left(\frac{\partial t_k}{\partial s}, \frac{\partial x_k}{\partial s}\right)$, defined by

$$\frac{\partial x_k}{\partial s} \Big/ \frac{\partial t_k}{\partial s} = \lambda_k$$

If the system of equations (1.1.2) can be reduced to such a canonical form (1.2.10), we say that the system is hyperbolic.

The ideas outlined above in one space-dimension can be generalized to M space-dimensions to define hyperbolicity in M dimensions (Garabedian [17], Jeffrey and Tanuiti [36]).

## 1.3 Systems of Conservation Laws: Physical Examples.

If we consider a perfect inviscid fluid in two space dimensions with horizontal and vertical components of velocity $u(x, y, t)$, $v(x, y, t)$ respectively, pressure $p(x, y, t)$ and density $\rho(x, y, t)$, then the Eulerian formulation of the partial differential equations for $u$, $v$, $p$ and $\rho$ express in mathematical language the physical laws of conservation of mass, momentum and energy.

Consider a quantity of the fluid contained in some plane region $D$ bounded by a simple closed curve $\delta D$, then the law of conservation of mass asserts that the rate at which fluid flows across $\delta D$ into $D$ must just balance the rate at which the total amount of fluid in $D$ increases with time.

In mathematical form this law is stated as

$$(1.3.1) \qquad \int_{\delta D} \left(\rho u \frac{\partial x}{\partial \nu} + \rho v \frac{\partial y}{\partial \nu}\right) ds = \frac{\partial}{\partial t} \iint_D \rho \, dxdy \quad ,$$

where $\frac{\partial}{\partial \nu}$ represents differentiation in the direction of the inner normal to $\delta D$ and $s$ represents the arc length measured along $\delta D$. By using the divergence theorem applied to the integral on the left hand side of (1.3.1) and simultaneously bringing the differential operator on the right hand side of (1.3.1) under the sign of integration, we obtain

$$\iint_D \left\{ \frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho u) + \frac{\partial}{\partial y}(\rho v) \right\} dxdy = 0,$$

from which we conclude, since $D$ can be taken as small as we please, that

$$(1.3.2) \qquad \frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho u) + \frac{\partial}{\partial y}(\rho v) = 0,$$

an equation which is said to be in <u>conservation form</u>. In general, the system of equations written in divergence form, namely,

$$(1.3.3) \qquad \frac{\partial w}{\partial t} + \sum_{i=1}^{M} \frac{\partial f_i}{\partial x_i} = 0 ,$$

is said to be in conservation form. (Here the $f_i$ are functions of the elements of $w$). Equation (1.3.3) expresses the fact that the quantity of $w$ contained in any $M$ dimensional hypervolume $V$ of x-space changes at a rate

$$\frac{d}{dt} \int_V w \, dx = \int_s f \cdot \nu \, ds,$$

where $s$ is the bounding surface of $V$ and $\nu$ is the inward drawn normal. Equation (1.3.2) is the well known equation of continuity.

Considering the momentum of the fluid in the x-direction when viscosity and all external forces such as gravity are neglected, we obtain the mathematical representation of the conservation of momentum, namely,

$$(1.3.4) \qquad \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} + \frac{1}{\rho} \frac{\partial p}{\partial x} = 0.$$

Similarly the conservation of momentum in the y-direction yields the differential equation

$$(1.3.5) \qquad \frac{\partial v}{\partial t} + u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} + \frac{1}{\rho}\frac{\partial p}{\partial y} = 0.$$

In order to formulate the law of conservation of energy as a partial differential equation in terms of u, v, p and $\rho$, it is necessary to introduce some concepts from thermodynamics. They involve the <u>entropy</u> S of the fluid and the <u>ratio of the specific heats $\gamma$</u> which is a constant exceeding unity. The hypothesis is made that the pressure p and the density $\rho$ of the fluid are connected with the entropy S by an equation of state of the form

$$(1.3.6) \qquad p = A(S)\rho^{\gamma}$$

where A(S) is positive but otherwise arbitrary. It can be shown (c.f. Courant-Friedrichs [9]) that energy is conserved if and only if the entropy S remains constant along each particle path, which means that

$$\frac{\partial S}{\partial t} + u\frac{\partial S}{\partial x} + v\frac{\partial S}{\partial y} = 0,$$

or using the equation of state (1.3.6),

$$(1.3.7) \qquad \frac{\partial}{\partial t}\left(\frac{p}{\rho^{\gamma}}\right) + u\frac{\partial}{\partial x}\left(\frac{p}{\rho^{\gamma}}\right) + v\frac{\partial}{\partial y}\left(\frac{p}{\rho^{\gamma}}\right) = 0.$$

Hence we have four equations formulating the conservation laws of mass, two components of momentum, and energy.

As can be seen, equation (1.3.2) falls into the category of conservation form defined above, for we can write $\rho u = f_1$ and $\rho v = f_2$ and obtain

$$\frac{\partial \rho}{\partial t} + \sum_{i=1}^{2} \frac{\partial f_i}{\partial x_i} = 0. \qquad \begin{array}{l} x_1 = x \\ x_2 = y \end{array}$$

whilst the equations (1.3.4), (1.3.5) and (1.3.7), as they stand, are not in conservation form. However, if we introduce the variable E defined by

$$p = (\gamma - 1)\left\{E - \tfrac{1}{2}\rho(u^2 + v^2)\right\}$$

the four equations (1.3.2), (1.3.4), (1.3.5) and (1.3.7) may be re-written in the vector form

$$\frac{\partial w}{\partial t} + \frac{\partial f}{\partial x} + \frac{\partial g}{\partial y} = 0$$

where

$$w = \begin{bmatrix} \rho \\ \rho u \\ \rho v \\ E \end{bmatrix}$$

$$f(w) = \begin{bmatrix} \rho u \\ -\frac{(\gamma - 3)}{2}\rho u^2 + (\gamma - 1)E - \frac{(\gamma - 1)}{2}\rho v^2 \\ \rho uv \\ \gamma E u + \frac{(\gamma - 1)}{2}(u^3 + v^2 u)\rho \end{bmatrix}$$

and $$g(w) = \begin{bmatrix} \rho v \\ \rho uv \\ -\frac{(\gamma - 3)}{2}\rho v^2 + (\gamma - 1)E - \frac{(\gamma - 1)}{2}\rho u^2 \\ \gamma E v + \frac{(\gamma - 1)}{2}(v^3 + vu^2)\rho \end{bmatrix}$$

This is now in conservation form.

Another example of conservation laws giving rise to partial differential equations in conservation form arises in considering the equations of magnetohydrodynamics which yield the Lundquist equations [47]. Problems involving such equations may be solved numerically by methods which are designed for systems in conservation form. However, systems do exist which cannot be written in this convenient form, or if they can, the process of transforming the equations into conservation form is so complex that the effort of transformation is not justified. It is more convenient to consider such equations in their stated non-conservation or quasilinear form and to treat them by methods designed for quasilinear systems (see (2.1.11)).

Finally in this section we briefly mention the concept of "Weak solutions". It is well known that solutions of the scalar equation

$$\frac{\partial u}{\partial t} + a(u)\frac{\partial u}{\partial x} = 0 \qquad (1.3.8)$$

subject to the initial condition $u = \phi(x)$, satisfy the relation

$$u = \phi(x - a(u)t)$$

A solution which satisfies this relation is termed a <u>soft solution</u> of (1.3.8). Because of the non-linearity of equation (1.3.8), in general a soft solution will not exist for all time. Instead we have to seek <u>weak solutions</u> defined by the requirement that

$$\iint (w_t u + f w_x)\,dx\,dt + \int w(x, o)\,\phi(x)\,dx = 0$$

where $\phi(x) = u(x, o)$ be satisfied for all smooth functions $w$ which vanish for $|x| + t$ sufficiently large. The concepts of soft and weak solutions are dealt with in [42] by Lax, in [50] by Hoh and Protter and by Jeffrey and Tanuiti [36].

Having introduced some physical examples which fall into the class of problems which will be considered, we discuss, in the next section, some methods available to us for the solution of these problems.

## 1.4 Methods of Solution.

In this section we shall consider briefly the methods available for solving systems of equations which may be of the form considered in the previous section. In considering the examples quoted there, we see that the complexity of these equations makes the possibility of obtaining exact solutions extremely remote. We consider therefore, approximate methods and mention three of them here:

(i) Approximate Analytic methods. Here a truncated series is usually employed to approximate the solution of the differential equation. There are limitations on these methods as the coefficients of the series are usually difficult to calculate and furthermore the series may not converge in the full region where the solution is required. However, this method can give extremely useful information regarding the behaviour of the solution in the areas where the series is convergent.

(ii) The method of characteristics. The basic principle involved in the use of the method of characteristics is that by an appropriate choice of co-ordinates, the original system of hyperbolic first order partial differential equations can be replaced by a system expressed in characteristic co-ordinates in terms of which integration is much simplified. Further, it can be shown that the reduction becomes particularly simple when applied to systems involving only two independent variables.

Consider the two first order simultaneous quasilinear equations

$$a_1 u_x + b_1 u_y + c_1 v_x + d_1 v_y = f_1 \tag{1.4.1}$$

(1.4.2) $$a_2u_x + b_2u_y + c_2v_x + d_2v_y = f_2$$

where $a_1$, $a_2$, ..., $f_1$, $f_2$ are functions of x, y, u and v. We may add to these the equations

(1.4.3) $$du = u_x dx + u_y dy$$

(1.4.4) $$dv = v_x dx + v_y dy$$

and the four equations may be written in vector form as

(1.4.5) $$\begin{bmatrix} a_1 & b_1 & c_1 & d_1 \\ a_2 & b_2 & c_2 & d_2 \\ dx & dy & 0 & 0 \\ 0 & 0 & dx & dy \end{bmatrix} \begin{bmatrix} u_x \\ u_y \\ v_x \\ v_y \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ du \\ dv \end{bmatrix}$$

The characteristic equations are obtained by equating to zero the determinant of the matrix of (1.4.5). This yields the quadratic in $\frac{dy}{dx}$, -

(1.4.6) $$(a_1c_2 - a_2c_1)(dy)^2 - (a_1d_2 - a_2d_1 + b_1c_2 - b_2c_1)\, dxdy + (b_1d_2 - b_2d_1)(dx)^2 = 0$$

Considering (1.4.5) again we see that in order for there to be any solution at all for the first derivatives when the determinant of the matrix vanishes, the right hand column vector must be compatible with this vanishing; that is, if we substitute the column vector in the matrix for one of its columns, the determinant of the resulting matrix must also vanish. For instance,

$$\begin{vmatrix} f_1 & b_1 & c_1 & d_1 \\ f_2 & b_2 & c_2 & d_2 \\ du & dy & o & o \\ dv & o & dx & dy \end{vmatrix} = 0,$$

which implies

$$(1.4.7) \quad [(b_1c_2 - b_2c_1)\frac{dy}{dx} - (b_1d_2 - b_2d_1)]du + [(d_1c_2 - d_2c_1)\frac{dy}{dx}]dv + $$

$$+ [(c_1f_2 - c_2f_1)\frac{dy}{dx} - (d_1f_2 - d_2f_1)]dy = 0$$

The method of characteristics consists of solving the two ordinary differential equations (1.4.6) in order to locate the characteristic curves and then to integrate (1.4.7) along these characteristic curves. In practice the calculations are carried out numerically since only very special problems can be integrated exactly by this method.

(iii) <u>The method of finite differences</u>, which forms the basis of the work presented in this thesis, is a method where differential operators are approximated by <u>difference operators</u>. The important property of the method is that it is universally applicable to both linear and non-linear problems.

The concepts and notations of the method of finite differences are discussed in the next section. Before proceeding to that section, some of the advantages of the method of characteristics over the method of finite differences will be discussed, together with some of the disadvantage

Lax [41] has shown that in an initial value problem, if the initial data is given as a function with continuous derivatives nearly everywhere, then the discontinuities in the initial data are propagated along the characteristics of the system. Therefore the method of characteristics treats discontinuities explicitly. On the other hand, the method of finite differences uses a particular form of differencing, or the method of pseudo-viscosity (Von Neumann and Richtmyer [69]) to represent the discontinuities The discontinuities are spread over two or three mesh intervals by these devices.

One of the disadvantages of the method of characteristics is that it is more complicated and, therefore, involves more programming effort, while the standard finite difference approach can solve both complex and simple problems with the same ease, because of the automatic handling of problems with discontinuities. Moreover, the generalization to many dimensions is relatively straight forward for the method of finite differences whereas even for the case of three independent variables the method of characteristics is extremely complex, although a great deal of work has been done recently along these lines by, for example, Sauerwein [60], Heie and Leigh [31].

## 1.5 Finite differences, notations and concepts.

In the method of finite differences, the region, in which the solution of the partial differential equation is required, is covered by a mesh or grid parallel to the principal axes. In the problems we shall consider these axes are always cartesian and the grid is hence rectangular, see figure (1.5.1).

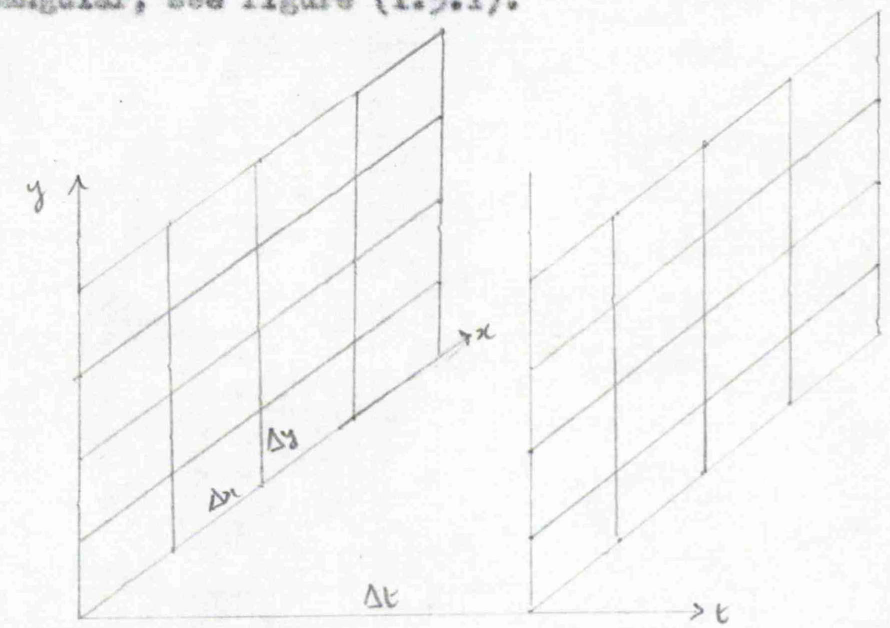


figure (1.5.1)

The mesh spacings are represented by $\Delta x$, $\Delta y$ and $\Delta t$ in the x, y and t directions. The quantities $\Delta x$ and $\Delta y$ will always be taken to be equal with $\Delta x = \Delta y = h$, and we represent $\Delta t$ by k. The points of intersection of the lines parallel to the principal axes are termed nodal, lattice or mesh points. It is at these mesh points that we determine approximate values of the solution of the differential equation.

It is convenient to represent difference methods in terms of difference operators defined by the relations

(1.5.1) $$\Delta_x u_{ij}^m = u_{i+1\, j}^m - u_{ij}^m$$ (the forward x - difference operator)

(1.5.2) $$\nabla_x u_{ij}^m = u_{ij}^m - u_{i\, j-1}$$ (the backward x - difference operato

$$\mu_x u_{ij}^m = \tfrac{1}{2}(u_{i+1\, j}^m + u_{i-1\, j}^m)$$ (the averaging operator)

where we have used the notation

$$u_{ij}^m = u(ih,\ jh,\ mk),$$

where i, j and m refer to a specific mesh point. Similar definitions apply to $\Delta_y$, $\nabla_y$, $\mu_y$, $\Delta_t$ and $\nabla_t$.

If u is assumed to be sufficiently differentiable, and we expand (1.5.1) and (1.5.2) using Taylor's series, then

(1.5.3) $$\Delta_x u_{ij}^m = \left(h \frac{\partial u}{\partial x} + \frac{h^2}{2!}\frac{\partial^2 u}{\partial x^2} + \frac{h^3}{3!}\frac{\partial^3 u}{\partial x^3} + \dots\right)_{ij}^m$$

and

(1.5.4) $$\nabla_x u_{ij}^m = \left(h \frac{\partial u}{\partial x} - \frac{h^2}{2!}\frac{\partial^2 u}{\partial x^2} + \frac{h^3}{3!}\frac{\partial^3 u}{\partial x^3} + \dots\right)_{ij}^m$$

respectively. Neglecting terms of order $h^2$ and higher, it is easily seen that

(1.5.5) $$\frac{1}{h}\Delta_x u_{ij}^m = \left(\frac{\partial u}{\partial x}\right)_{ij}^m + O(h),$$

and

(1.5.6) $$\frac{1}{h}\nabla_x u_{ij}^m = \left(\frac{\partial u}{\partial x}\right)_{ij}^m + O(h),$$

are both approximations to the first partial derivative $\frac{\partial u}{\partial x}$ and the error is first order in h (these formulae will later be said to have second order truncation error).

If (1.5.3) and (1.5.4) are added, a more accurate approximation to $\left(\frac{\partial u}{\partial x}\right)_{ij}^{m}$ (with third order truncation error) is obtained, namely

(1.5.7) $$\frac{1}{2h}(\Delta_x + \nabla_x)u_{ij}^{m} = \left(\frac{\partial u}{\partial x}\right)_{ij}^{m} + O(h^2)$$

The left hand side of (1.5.7) is termed the central difference approximatio to $\frac{\partial u}{\partial x}$.

If a partial differential equation

(1.5.8) $$Du = 0$$

is approximated by a difference equation

(1.5.9) $$D_{h,k}u = 0,$$

then the difference between (1.5.8) and (1.5.9)

(1.5.10) $$Du - D_{h,k}u$$

is termed the discretization error. The order of accuracy is given by the lowest order term in the difference (1.5.10). Consider D to be given by

$$Du = \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0 \qquad (1.5.11)$$

Consider $D_{h,k}u$ to be given first by

$$D_{h,k}u = \frac{1}{k}[u_i^{m+1} - u_i^m + p/2(u_{i+1}^m - u_{i-1}^m)] \qquad (1.5.12)$$

where $p = k/h$ is the mesh ratio. Expanding the right hand side of (1.5.12) using Taylor's series and consequently determining the difference (1.5.10), it can be shown easily that the order of accuracy is $O(k + h^2)$.

If instead $D_{h,k}u$ is given by

$$D_{h,k}u = \frac{1}{k}[u_i^{m+1} - \mu_x u_i^m + p/2(u_{i+1}^m - u_{i-1}^m)] \qquad (1.5.13)$$

we obtain a second difference approximation to (1.5.11) which, in a similar way to that described above, may be shown to have an accuracy again of $O(k + h^2)$. Hence on considerations of accuracy, it is not clear which method is the more desirable in an actual calculation. This question is answered on the basis of the stability of the finite difference method. The concept of stability is studied later in this section.

Finite difference approximations are either explicit or implicit. An explicit formula involves only one mesh point at the advanced time level. Implicit procedures involve more than one mesh point at the advanced time level and require the solution of sets of equations at each time level.

Before going on to consider the concepts of <u>convergence</u> and <u>stability</u>, let us consider the following interesting example of a finite difference method for the wave equation in one space dimension, namely

$$u_{tt} = u_{xx} \tag{1.5.14}$$

subject to the initial data

$$u(x, 0) = f(x), \qquad u_t(x, 0) = g(x) \tag{1.5.15}$$

where f and g are functions possessing several continuous derivatives. To formulate a finite difference approximation to this initial value problem we replace (1.5.14) by

$$\frac{u_i^{m+1} - 2u_i^m + u_i^{m-1}}{k^2} = \frac{u_{i+1}^m - 2u_i^m + u_{i-1}^m}{h^2} \tag{1.5.16}$$

and replace (1.5.15) by the discrete conditions

$$u_i^0 = f_i \quad \text{and} \quad u_i^1 = f_i + kg_i + \frac{k^2}{2} \frac{\partial^2}{\partial x^2} f_i \,, \tag{1.5.17}$$

h and k being the mesh spacings in the space and time directions respectively. Assume that f and g have been prescribed over an interval $0 \leq x \leq 2Nh$ of the x-axis so that $f_i$ and $g_i$ are known in the corresponding range $0 \leq i \leq 2N$. Then the initial conditions (1.5.17) define

$u(ih, 0) = u_i^0$ and $u(ih, k) = u_i^1$ for $0 \leq i \leq 2N$ and it is natural to try to determine $u_i^m$ for larger choices of the index $m$ by applying the difference equation (1.5.16). This may be effected by rearranging (1.5.16) to the form

$$(1.5.18) \qquad u_i^{m+1} = 2u_i^m - u_i^{m-1} + \alpha(u_{i+1}^m - 2u_i^m + u_{i-1}^m)$$

where $\alpha = k^2/h^2$, which yields the unknown $u_i^{m+1}$ at the level $(m + 1)k$ in terms of values at the previous two levels $mk$ and $(m - 1)k$. (We note here that as $m$ increases, the number of unknowns that may be determined reduces at each time step owing to the increasing lack of information at either end of the two known lines of data, so that eventually only <u>one</u> point value will be determined).

If however $\alpha$ is chosen to be unity, the difference scheme (1.5.18) reduces to the simpler form

$$(1.5.19) \qquad u_i^{m+1} = u_{i+1}^m - u_i^{m-1} + u_{i-1}^m$$

from which a closed expression can be found for the solution to the initial value problem (see, for example, Garabedian [17], page 466). The exact solution involves summation along a pair of characteristics of the wave equation and so solving (1.5.19) is equivalent to solving the wave equation by the method of characteristics.

In using finite difference methods, one has to consider the concepts of <u>convergence</u> and <u>stability</u> of the difference schemes; that is, under what conditions does the solution $u_{ij}^m$ of the difference equation tend to the solution u(x, y, t) of the differential equation as the mesh sizes tend to zero, and under what conditions do the errors that accumulate in the numerical computation of $u_{ij}^m$ remain bounded for an increasing number of time steps.

This thesis will concern the numerical solution of <u>non-linear</u> systems of partial differential equations and it is therefore necessary to have some criteria for the convergence and stability of the methods used. The stability for linear problems, and for <u>some</u> problems with variable coefficients (Lax [43], Strang [62]) has been studied to a considerable extent. The stability for non-linear problems is far more complex. It is assumed that the stability of a difference scheme for a non-linear proble depends upon the stability of the first variation of the difference operator, that is, on the linearized difference equation. That this assumption is justified (at least when the partial differential equation and the true solution possess a certain degree of smoothness) is the subject of a paper by Strang [62].

Before giving the theorem of Strang we define the $L_2$ norm. Consider some function u(x) which is a well behaved function in some region X; then we define the inner product as

$$(u, v) = \sum_j u(x_j)\, v^*(x_j)$$

where $v^*$ is the complex conjugate transpose of $v$. Then the $L_2$ norm of the function u(x) is defined by

$$\| u \|_2 = \{ (u, u) \}^{\frac{1}{2}}.$$

This definition may be extended in the obvious way to the $L_2$ norm defined on difference operators when

$$\| D \|_2 = \sup_{\|u\| = 1} \| Du \|_2 .$$

The theorem of Strang asserts the following. Consider a consistent difference approximation to a problem in which the unknown vector and Jacobian matrix (the matrix A of equation (1.2.1)) have several continuous derivatives. Then, in order that convergence of the difference scheme follow, the first variation of the difference operator should be stable in the $L_2$ norm. Hence we consider the stability of the linear problem.

Consider a two-level difference scheme for a two dimensional problem written in the form

$$(1.5.20) \quad \sum_{i,j} \tilde{A}_1^{i,j} u_{m+1}(x + ih, y + jh) = \sum_{i,j} \tilde{A}_2^{i,j} u_m(x + ih, y + jh)$$

where $\tilde{A}_r^{i,j}$ $r = 1, 2$ are coefficient matrices, and the summations are not necessarily the same at the two time levels. Introducing the notation of difference operators from (1.5.1) and (1.5.2), (1.5.20) may be written in the form

$$(1.5.21) \quad A_1(\Delta_x, \nabla_x, \Delta_y, \nabla_y)u_{m+1} = A_2(\Delta_x, \nabla_x, \Delta_y, \nabla_y)\, u_m$$

where $A_1$, $A_2$ are difference operators. Assuming the existence of $A_1^{-1}$, then (1.5.21) may be written as

$$u_{m+1} = Cu_m$$

where C may involve an infinite expansion in terms of the difference operators. On substituting Fourier series for $u_{m+1}$ and $u_m$ in (1.5.20), followed by the cancellation of a common factor (see the example below) the remaining terms may be written as

$$G_1 V_{m+1} = G_2 V_m$$

where $G_j = \sum_{r,s} \tilde{A}_j^{r,s} \exp\{i(\beta rh + \gamma sh)\} \qquad j = 1, 2$

and $\beta$, $\gamma$ are real numbers. Thus

$$V_{m+1} = G(\beta, \gamma) V_m,$$

where $G(\beta, \gamma)$ is termed the amplification matrix.

Stability Analysis is concerned with the behaviour of the amplification matrix. First there is the necessary condition for stability due to Von Neumann [51], namely

$$|\lambda_i| \leq 1 + O(k) \qquad i = 1, 2, \ldots n$$

where $\lambda_i$ is an eigenvalue of G. The O(k) term is included to account for the possible genuine growth of the solution to the problem with time. Second, the sufficient condition for stability due to Lax-Richtmyer [44] and Kreiss [38], [40], is

$$|\mu_i|^{\frac{1}{2}} \leq 1 + O(k) \qquad i = 1, 2, \ldots n$$

where $\mu_i$ is an eigenvalue of $G^*G$ where $G^*$ is the hermitian transpose of G.

Having given these two conditions for stability, we now indicate how they may be applied to non-linear problems. Consider the non-linear partial differential equation

$$\frac{\partial u}{\partial t} + \frac{\partial f}{\partial x} = 0 \qquad \text{(u and f vectors)}$$

and the finite difference method

$$(1.5.22) \qquad u_i^{m+1} = u_i^m - p\Delta_x f_i^m.$$

In order to investigate the conditions under which this method would be stable, $\Delta_x f_i^m$ is linearized, that is

$$\Delta_x f_i^m \rightarrow A\Delta_x f_i^m$$

where A is the Jacobian matrix of f with respect to the components of u. The matrix A is then assumed (locally) constant and the equation (1.5.22) becomes

$$u_i^{m+1} = u_i^m - Ap\Delta_x u_i^m.$$

The stability analysis is then performed by considering the eigenvalues of the amplification matrix as indicated above.

To illustrate the above stability theory consider the following example of a first order scalar hyperbolic equation

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0$$

approximated by the difference scheme

$$u_i^{m+1} = \mu_x u_i^m - \frac{p}{2} H_x u_i^m.$$

where $H_x \equiv (\Delta_x + \nabla_x)$. Introducing the Fourier transform, we obtain

$$e^{i\beta x} e^{\lambda t} e^{\lambda k} = \tfrac{1}{2}(e^{i\beta(x+h)} + e^{i\beta(x-h)})e^{\lambda t} - \frac{p}{2}(e^{i\beta(x+h)} - e^{i\beta(x-h)})e^{\lambda t}$$

Cancelling the common factor $e^{\lambda t} e^{i\beta x}$ then

$$e^{\lambda k} = \tfrac{1}{2}(e^{i\beta h} + e^{-i\beta h}) - \tfrac{1}{2}p(e^{i\beta h} - e^{-i\beta h})$$

giving

$$e^{\lambda k} = \cos \beta h - ip \sin \beta h ,$$

so that

$$|e^{\lambda k}| = (1 + (p^2 - 1) \sin^2 \beta h)^{\frac{1}{2}}$$

From which the method is stable if $p^2 < 1$.

However, the method

$$u_i^{m+1} = u_i^m - \frac{p}{2} H_x u_i^m$$

gives

$$e^{\lambda k} = 1 - ip \sin \beta h$$

the modulus of which is always greater than 1 and the method can only be stable in certain circumstances.

We next discuss the concept of convergence. Consider the finite difference analogue

$$(1.5.23) \qquad u_{m+1} = B(k, h)u_m$$

of the differential problem consisting of

$$(1.5.24) \qquad u_t = Au \qquad\qquad u(0) = u_0,$$

where $u_0$ represents a preassigned initial state of the system and $A$ is a differential operator. This formulation is not restricted to explicit difference schemes. If the system is implicit, the operator $B$ will contain the inverse of a difference operator.

The concept of convergence presupposes an infinite sequence of calculations with increasingly finer mesh. If we assume that $h = g(k)$ defines the manner in which $h \rightarrow 0$ as $k \rightarrow 0$, we may write (1.5.23) as

$$(1.5.25) \qquad u_{m+1} = C(k)u_m.$$

Equation (1.5.25) is called a <u>consistent</u> approximation to the differential equation (1.5.24) if

$$\lim_{k \rightarrow 0} \left\| \left( \frac{C(k) - I}{k} - A \right) u(t) \right\| = 0 \text{ uniformly in } t \text{ for } 0 \leq t < T.$$

This simply requires that the discretization error of the difference scheme tends to zero as h and k tend to zero. If we define the operator E(t) such that

$$u(t) = E(t)u_o,$$

then we say C(k) is a <u>convergent</u> approximation to the differential problem (1.5.24) if for any sequence $K_j$, $M_j$ such that $K_j$ tends to zero and $M_jK_j \rightarrow t$ for $o \leq t \leq T$, then

$$\left\| \left\{ C(K_j) \right\}^{M_j} u_o - E(t)u_o \right\| \rightarrow o.$$

Finally we mention briefly the substance of the (now classical) Courant-Friedrichs-Lewy condition [10] for convergence and stability of (explicit) finite difference schemes. This states that the <u>domain of dependence</u> of the differential scheme must be contained within the domain of dependence of the difference scheme. This is best described with the aid of figures (1.5.2), (1.5.3).

- characteristics of differential problem

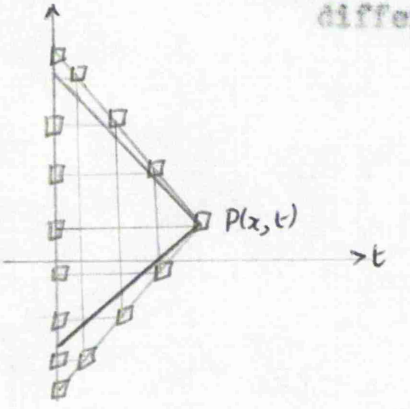


figure (1.5.2)

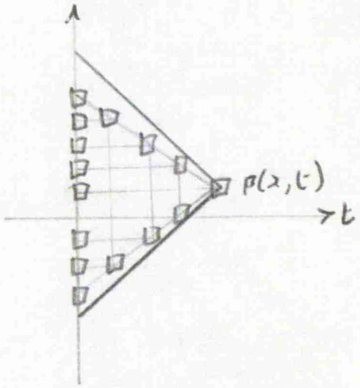


figure (1.5.3)

Consider the wave equation in one space dimension, namely

$$u_{tt} = u_{xx}, \tag{1.5.26}$$

when the characteristic directions are found to be (see section (1.2))

$$\frac{dx}{dt} = \pm 1$$

Thus if we consider the solution of (1.5.26) at a point P(x, t) in <u>figure (1.5.2)</u> by a difference method which determines the solution at P in terms of the nodal values marked □, then it can clearly be seen that the mesh points on the initial line which are used for the solution at P, contain points outside those lying between the points of intersection of the characteristics passing through P with gradients ± 1. In this case, the method will converge, Courant-Friedrichs-Lewy [10]. However, considering the solution of (1.5.26) at a point P(x, t) in <u>figure (1.5.3)</u> by a difference method which determines the solution at P in terms of the nodal values marked □, then it can clearly be seen that there are points on the initial line lying within the points of intersection of the characteristics passing through P with gradients ± 1 which are not used by the difference scheme and hence the difference method will not converge.

## 1.6 Survey of existing methods. Outline of the thesis.

In recent times, various approximations have been given for particular non-linear problems. For systems of conservation laws similar to those introduced in section (1.3), finite difference schemes have been proposed by Friedrichs [14], Longley [46], Fromm [16], [illegible] and Lax and Wendroff [45]. The scheme of Friedrichs [14] is given by

$$(1.6.1) \qquad w_i^{m+1} = \mu_x w_i^m - p/2 \, H_x f_i^m$$

for the first order system of conservation laws

$$(1.6.2) \qquad w_t + f_x = 0$$

A scheme commonly called the 'leap-frog' scheme was recently the subject of investigation in an experimental paper by Richtmyer and Morton [59] for the solution of equation (1.6.2). This scheme is a **three-level** scheme and may be represented by

$$(1.6.3) \qquad w_i^{m+1} = w_i^{m-1} - pH_x f_i^m .$$

The method of Lax and Wendroff [45] for (1.6.2) may be written as

$$(1.6.4) \qquad w_i^{m+1} = w_i^m - \frac{p}{2} H_x f_i^m + \frac{p^2}{2} \delta_x A^2 \delta_x w_i^m$$

where A is the Jacobian matrix of f, and $\delta_x$ is the difference operator defined by

$$\delta_x u_i^m = u_{i+\frac{1}{2}}^m - u_{i-\frac{1}{2}}^m$$

**This method, (1.6.4), can be written as a <u>two-step</u> method (Richtanger [58]):**

$$(1.6.5) \qquad w_i^{m+1} = \mu_x w_i^m - \frac{p}{2} H_x f_i^m$$

$$(1.6.6) \qquad w_i^{m+2} = w_i^m - pH_x f_i^{m+1}$$

We note that (1.6.5) followed by (1.6.6) is nothing more than Friedrich's scheme (1.6.1) followed at the next step by the leap-frog scheme (1.6.3). Whilst some specialized methods do exist, for example for the equations of meteorology, Arakawa [2], Marchuk [48], and others, most calculations have been performed using the Lax-Wendroff method (1.6.4), Richtmyer's two step version [58], or variations of these methods. For example, Burstein [5], [6] has used variations of the above methods for calculating fluid dynamical problems with shocks.

In chapter 2 we will consider generalizations of (1.6.4) and (1.6.5, 1.6.6) applied to the system of conservation laws (1.6.3) and to their two dimensional counterparts, namely

$$(1.6.7) \qquad \frac{\partial u}{\partial t} + \frac{\partial f}{\partial x} + \frac{\partial g}{\partial y} = 0$$

The corresponding nonconservative equations in one and two dimensions are considered and the corresponding generalizations derived. It is found that the explicit methods require boundary data which is not given in the differential problem. In chapter 3, techniques are discussed for the introduction of extra data required by the explicit schemes. Chapter 4 contains the above procedures applied to the more general problems

$$\frac{\partial u}{\partial t} + \frac{\partial f}{\partial x}(u, x, t) = \gamma(u, x, t)$$

and

$$\frac{\partial u}{\partial t} + A(u, x, t)\,\frac{\partial u}{\partial x} = \gamma(u, x, t)$$

and their corresponding two dimensional analogues. In chapters 2 and 4, implicit methods, including alternating direction implicit methods (see later) are also derived. The thesis is concluded in Chapter 5 with a brief consideration of the behaviour of the methods applied to one dimensional problems with discontinuities in their initial data, together with a new formulation of a scheme introduced by Strang [61], [62] for two dimensional problems.

# CHAPTER II

# THE GENERAL PREDICTOR - CORRECTOR METHOD

Gourlay and Morris [24]

## Introduction

In deriving accurate finite difference methods for non-linear partial differential equations, one often finds that the finite difference method is also non-linear. For example, the Crank-Nicholson method for the equation

$$\frac{\partial u}{\partial t} + \frac{\partial f}{\partial x} = 0$$

would be

$$u_i^{m+1} = u_i^m - {}^{p}/_{4}H_x(f_i^{m+1} + f_i^m)$$

where the notation of section (1·5) has been used and the difference operator $H_x$ is defined by the relation

$$H_x u_i^m = (\Delta_x + \nabla_x)u_i^m = u_{i+1}^m - u_{i-1}^m$$

As can be seen, the above difference formula involves the non-linear term $f_i^{m+1}$ on the right hand side and hence requires a recursive method in order to determine $u_i^{m+1}$. A method by which this difficulty is overcome is the "predictor - corrector" method (similar to that derived in the non-linear ordinary differential equations, see for example [55] ) by which a first formula is used to approximate the differential equation and then a second formula, using the solution to the first is used to gain the overall accuracy of the method. Each formula yields the unknown linearly in terms of previously calculated functions.

In this chapter, predictor - corrector schemes are derived which involve a variable parameter, a special value of which yields the two step Lax-Wendroff method [58]. Explicit and implicit methods are derived for both the one dimensional equations (1·6·2) and the two dimensional equations (1.6.7). The two dimensional implicit method uses a technique of solution designated the A.D.I. (Alternating Direction Implicit) method. An A.D.I. method is one whereby a partial difference equation in several space variables is solved by a multi-step finite difference procedure of a particular type. The name arises from the fact that at each step in the procedure we solve along lines parallel to a space co-ordinate axis, the axis being different at each step. Although the finite difference scheme at each stage is implicit in nature, it is of a particular form which may be solved by a direct or non-iterative method.

The contents of this chapter extend and generalise some of the work presented in the final chapter of the doctoral thesis of Gourlay [19].

## 2•1 The Explicit One Dimensional Case.

Richtmyer in his review paper [58] considers the Lax-Wendroff method [45] for the partial differential equation

$$(2\cdot1\cdot1) \qquad \frac{\partial u}{\partial t} + \frac{\partial f}{\partial x} = 0$$

where u is an n-column unknown vector and f is a function of u as a two-step process, namely

$$(2\cdot1\cdot2) \qquad u_i^{*m+1} = \mu_x u_i^m - {}^p/_4 H_x f_i^m$$

$$(2\cdot1\cdot3) \qquad u_i^{m+1} = u_i^m - {}^p/_2 H_x f_i^{*m+1}$$

where $f_i^{*m+1} = f(u_i^{*m+1})$. Equation (2•1•2) is the predictor formula referred to in the introduction of this chapter and is a first order correct approximation to $u(ih, (m + \frac{1}{2})k)$. The overall scheme is second order correct.

The question naturally arises; is the positioning of the predictor level 'best', in some sense? Could not some other position be found, for instance, at $(m + 1)k$ or even in advance of that whereby better results may be obtained? We investigate the existence of methods with such positioning of predictor levels in what follows.
If in (2•1•2) a variable parameter 'a', say, is introduced, the resulting equation may be written as

$$(2\cdot1\cdot4) \qquad u_i^{*m+1} = \mu_x u_i^m - apH_x f_i^m$$

and $u_i^{*m+1}$ is a first order approximation to $u(ih, (m+2a)k)$.

The equation

$$(2\cdot1\cdot5) \qquad u_i^{m+1} = u_i^m - p(bH_x f_i^m + cH_x f_i^{*m+1})$$

is introduced instead of (2·1·3), where b and c are constants which are to be determined.

If the * level is eliminated between (2·1·4) and (2·1·5), and the resulting expression expanded about $u_i^m$ using Taylor's theorem and retaining terms up to and including those of order $h^2$, we obtain

$$(2\cdot1\cdot6) \quad u_{m+1} = u_m - 2(b + c)k\frac{\partial f}{\partial x}_m + 4ack^2 \frac{\partial}{\partial x}\left(A\frac{\partial f}{\partial x}\right)_m + O(k^3)$$

where we have used the abbreviated notation

$$u_{m+1} = u_i^{m+1} \quad \text{and} \quad u_m = u_i^m$$

A Taylor expansion of $u_{m+1}$ in terms of $u_m$ and its derivatives yields

$$(2\cdot1\cdot7) \quad u_{m+1} = \left(u + k\frac{\partial u}{\partial t} + \frac{k^2}{2}\,\frac{\partial^2 u}{\partial t^2} + \ldots\right)_m.$$

Using the differential equation (2·1·1), it is seen that

$$(2\cdot1\cdot8) \quad \frac{\partial^2 u}{\partial t^2} = -\frac{\partial^2 f}{\partial x \partial t} = \frac{\partial}{\partial x}\left(-\frac{\partial f}{\partial t}\right) = \frac{\partial}{\partial x}\left(-\frac{\partial f}{\partial u}\frac{\partial u}{\partial t}\right) = \frac{\partial}{\partial x}\left(A\frac{\partial f}{\partial x}\right)$$

Substituting $\frac{\partial^2 u}{\partial t^2}$ from (2·1·8) into (2·1·7) gives

$$(2\cdot1\cdot9) \quad u_{m+1} = \left(u - k\frac{\partial f}{\partial x} + \frac{k^2}{2}\frac{\partial}{\partial x}\left(A\frac{\partial f}{\partial x}\right)\right)_m + O(h^3)$$

Comparing coefficients between (2·1·6) and (2·1·9) it is seen in order that (2·1·4) and (2·1·5) be a second order correct approximation to (2·1·1) then

$$2(b + c) = 1$$

$$8ac = 1$$

must be satisfied. This set of equations has the solution

$$b = \tfrac{1}{2}\left(1 - \frac{1}{4a}\right)$$

$$c = \frac{1}{8a}$$

in terms of the parameter a. Formulae (2·1·4) and (2·1·5) now form the explicit two step method

$$(2\cdot1\cdot10)\qquad \begin{aligned} \overset{*}{u}_{m+1} &= \mu_x u^m_i - apH_x f^m_i \\ u_{m+1} &= u_m - {}^{p}/_{2} \left[\left(1 - \frac{1}{4a}\right)H_x f^m_i + \frac{1}{4a} H_x f^{*\,m+1}_i\right] \end{aligned}$$

It can be seen that by introducing $a = \frac{1}{4}$ into (2·1·10), the resulting scheme is the Lax-Wendroff method (2·1·2), (2·1·3). In a similar manner it may be shown that the corresponding method for the system of equations in non-conservation form, namely

$$\frac{\partial u}{\partial t} + A\frac{\partial u}{\partial x} = 0$$

where $A = A(u)$ is a function $u$, is

$$(2\cdot1\cdot11)\qquad \begin{aligned} \overset{*}{u}_{m+1} &= \mu_x u^m_i - apA_m H_x u^m_i \\ u_{m+1} &= u_m - {}^{p}/_{2}\left[\left(1 - \frac{1}{4a}\right)A_m H_x u^m_i + \frac{1}{4a} \overset{*}{A}_{m+1} H_x u^{*\,m+1}_i\right] \end{aligned}$$

where $A_m = A(u_m)$ and $\overset{*}{A}_{m+1} = A(\overset{*}{u}_{m+1})$.

## 2·2 The one dimensional implicit case.

In [18], Gary considered implicit predictor-corrector schemes for the numerical solution of problems in fluid dynamics. These methods were modifications of the basic procedure

$$(2{\cdot}2{\cdot}1)\qquad \begin{aligned} &\overset{*}{u}_{m+1} = u_m - {}^{p}/_{2}\, A_m H_x u_i^m \\ &u_{m+1} + {}^{p}/_{4}\, A\left(\frac{u_m + \overset{*}{u}_{m+1}}{2}\right) H_x u_i^{m+1} = u_m - {}^{p}/_{4}\, A\left(\frac{u_m + \overset{*}{u}_{m+1}}{2}\right) H_x u_i^m \end{aligned}$$

for the system of equations in non-conservation form

$$(2{\cdot}2{\cdot}2)\qquad \frac{\partial u}{\partial t} + A\frac{\partial u}{\partial x} = 0,$$

where $u$ is an unknown n-column vector and $A$ is an $n \times n$ matrix of the components of $u$ and we have used the notation $A_m = A(u_m)$.

An implicit predictor-corrector scheme will now be proposed which approximates the system of Conservation laws

$$(2{\cdot}2{\cdot}3)\qquad \frac{\partial u}{\partial t} + \frac{\partial f}{\partial x} = 0$$

where $u$ is again an unknown n-column vector and $f$ is a known function of $u$, and from which the form of the corresponding method for (2·2·2) may be derived. Consider the predictor-corrector scheme

$$(2{\cdot}2{\cdot}4)\qquad \overset{*}{u}_{m+1} = \mu_x u_i^m - apH_x f_i^m$$

$$(2{\cdot}2{\cdot}5)\qquad u_{m+1} = u_m - p[\,bH_x f_i^m + cH_x \overset{*}{\tilde{A}}_{m+1}\, u_{m+1}]$$

where $\tilde{A}$ is defined by the relation

$$\tilde{A}(u).u = f(u)$$

and

$$\overset{*}{\tilde{A}}_{m+1} = \tilde{A}(\overset{*}{u}_{m+1})$$

and a, b and c are constants to be determined. Eliminating the * level between (2·2·4) and (2·2·5), expanding the resulting expression about $u_i^m$ using Taylor's series and equating the coefficients of the resulting expression with those of the expansion of $u_i^{m+1}$ in terms of $u_i^m$ and its derivatives, it may easily be shown that

$$a = \tfrac{1}{2},\ b = c = \frac{1}{4}$$

must be satisfied in order that (2·2·4), (2·2·5) be a second order approximation to (2·2·3).

The implicit scheme is, therefore

(2·2·6) $\overset{*}{u}_{m+1} = \mu_x u_i^m - {}^{P}/_{2}\, H_x f_i^m$

(2·2·7) $u_{m+1} = u_m - {}^{P}/_{4}[H_x f_i^m + H_x \overset{*}{\tilde{A}}_{m+1}\, u_{m+1}]$.

Equation (2·2·7) may be written in the slightly different form as

(2·2·8) $(I + {}^{P}/_{4}\, H_x\, \overset{*}{\tilde{A}}_{m+1})u_{m+1} = u_m - {}^{P}/_{4}\, H_x f_i^m$,

where $I$ is the $n \times n$ unit matrix; $u_{m+1}$ may be obtained from (2·2·6) and (2·2·8) by a block tridiagonal matrix inversion - see, for example, Varga [68].

The corresponding implicit difference scheme for the system in non-conservation form, namely (2·2·2) is

(2·2·9) $\overset{*}{u}_{m+1} = \mu_x u_i^m - {}^{P}/_{2}\, A_m H_x u_i^m$

(2·2·10) $u_{m+1} = u_m - {}^{P}/_{4}\,[A_m H_x u_i^m + \overset{*}{A}_{m+1} H_x u_i^{m+1}]$

which may again be solved for $u_{m+1}$ by using block tridiagonal inversion techniques.

The use of the implicit schemes introduced above over the use of the method of Gary may be justified on the following grounds:- the predictor (2·2·4) has been used instead of the predictor used by Gary because, as can easily be verified, (2·2·4) is stable in the linearized sense whereas the predictor of (2·2·1) is an unstable one and, as the use of predictor-corrector methods in the solution of first order ordinary differential equations of the form $y' = f(x,y)$ has shown, the corrector must be iterated when the predictor is unstable if the overall scheme is to remain stable (see, for example [29]). Also, the scheme introduced by Gary was designed specifically for systems in non-conservation form, whereas the above implicit scheme is cast in the two required forms for solution of either conservation or non-conservation equations.

## 2·3 Stability and truncation error of the methods introduced in 2·1 and 2·2.

As we saw in (1·5), the stability of the finite difference approximations to non-linear partial differential equations is governed by the local amplification matrix.

We consider first the stability of the explicit predictor-corrector scheme (2·1·10). Linearizing (2·1·10) in a manner similar to that shown in section (1·5) we obtain

$$(2\cdot3\cdot1)\quad \begin{aligned} \overset{*}{u}_{m+1} &= \mu_x u_m - apA\, H_x u_m \\ u_{m+1} &= u_m - {}^{p}/_{2}\, AH_x\left[\left(1-\frac{1}{4a}\right)u_m + \frac{1}{4a}\,\overset{*}{u}_{m+1}\right] \end{aligned}$$

where the matrix $A$ is assumed to be locally constant and symmetric. Elimination of $\overset{*}{u}_{m+1}$ from the predictor-corrector method yields the result

$$(2\cdot3\cdot2)\quad u_{m+1} = u_m - {}^{p}/_{2}\, AH_x\left[u_m + \frac{1}{8a}\left(u^m_{i+1} - 2u^m_i + u^m_{i-1}\right)\right] + \frac{p^2A^2}{8} H_x^2 u^m_i.$$

If a Fourier decomposition of the errors is made in the usual way, it follows that the amplification matrix $G$ of (2·3·2) (and hence locally of (2·3·1)) is given by

$$G = I - \tfrac{1}{2}p^2A^2 \sin^2\beta h - pA\sqrt{-1}\,\sin\beta h\left[1 + \frac{1}{4a}(\cos\beta h - 1)\right]$$

where $\beta$ is a real number, $I$ is the $(n \times n)$ unit matrix. In section (1·5) we saw that the Lax-Richtmyer sufficient condition for stability requires

$$(2\cdot3\cdot3)\quad \|G\| \leq 1 + O(k)$$

where $\|\cdot\|$ represents the spectral norm. Equation (2·3·3) is satisfied if the eigenvalues of $G$ are less than one in modulus.

If the eigenvalues of $A$ are given by the roots of

$$|A - \lambda I| = 0,$$

it is easily seen that the eigenvalues $g$, $(g \geq 0)$ of $G$ are given by

$$(2\cdot3\cdot4) \quad |g| = 1 - p^2\lambda^2 \sin^2\beta h \left[1 - \frac{p^2\lambda^2}{4}\sin^2\beta h - \left(1 - \frac{1}{2a}\sin^2\beta h\right)^2\right]$$

and therefore for stability we require

$$(2\cdot3\cdot5) \quad 1 - \frac{p^2\lambda^2}{4}\sin^2\beta h + \left(1 - \frac{1}{2a}\sin^2\beta h\right)^2 \geq 0,$$

that is

$$(2\cdot3\cdot6) \quad p^2\lambda^2 \leq \frac{4a - \sin^2\frac{\beta h}{2}}{4a^2\cos^2\frac{\beta h}{2}}$$

for all real $\beta$. Since $p^2\lambda^2 \geq 0$, we have immediately that

$$4a - \sin^2\frac{\beta h}{2} \geq 0$$

for all real $\beta$. This requires that $a \geq \frac{1}{4}$. Rewriting the right hand side of (2·3·6) as

$$\frac{4a - 1 + \cos^2\frac{\beta h}{2}}{4a^2\cos^2\frac{\beta h}{2}},$$

and letting $s = \cos^2\frac{\beta h}{2}$ and

$$Q(s) = \frac{4a - 1 + s}{4a^2 s}$$

then

$$\frac{dQ}{ds} = \frac{1 - 4a}{4a^2 s^2}.$$

Since $Q(s) \geq 0$ and, from above, $a \geq \frac{1}{4}$, it follows that $\frac{dQ}{ds} \leq 0$ and that the minimum of $Q(s)$ occurs at $s = 1$.

It therefore follows for stability that

$$p^2\lambda^2 \leq \min_{0 \leq \cos^2\frac{\beta h}{2} \leq 1} \frac{4a - 1 + \cos^2\frac{\beta h}{2}}{4a^2\cos^2\frac{\beta h}{2}} = \frac{1}{a}$$

that is

$$p|\lambda| \leq \frac{1}{\sqrt{a}} \text{ for stability.}$$

Since the predictor corrector scheme is a two step method with the value of $u_i^{n+1}$ given ultimately in terms of $u_i^n$, $u_{i\pm1}^n$ and $u_{i\pm2}^n$ (that is after substituting the predictor level into the corrector level), the Courant-Friedrichs-Lewy condition [10] requires that $p|\lambda| \leq 2$. It therefore follows that we have "maximum" stability for $a = \frac{1}{4}$, and conditional stability for $p|\lambda| \leq \frac{1}{\sqrt{a}}$ for other values of $a > \frac{1}{4}$. A similar analysis obviously holds for the predictor-corrector scheme derived for the equations in non-conservation form.

An analysis of the implicit methods of section (2·2) may be undertaken in a similar way. The result of this analysis is that the implicit methods are stable in the linearised sense for all values of $p$. As stated in section (2·1) the explicit predictor-corrector scheme is an $0(h^2)$ accurate method. For completeness, the actual form of the principal part of the truncation error will be quoted and is

$$-ph^3\left[\left(\frac{a}{2}-\frac{1}{3!}\right)p^2\frac{\partial}{\partial x}\left\{\left(\frac{\partial f}{\partial x}\right)^2\frac{\partial^2 f}{\partial u^2}\right\}+\frac{1}{6a}\frac{\partial}{\partial x}\left\{\frac{\partial^2 u}{\partial x^2}\frac{\partial f}{\partial u}\right\}-\frac{p^2}{3!}\frac{\partial}{\partial x}\right.$$
$$\left.\left\{\frac{\partial f}{\partial u}\frac{\partial}{\partial x}\left(\frac{\partial f}{\partial u}\frac{\partial f}{\partial x}\right)\right\}+\frac{2}{3!}\frac{\partial^3 f}{\partial x^3}\right]$$

which is a function of the parameter a.

## 2·4 Two dimensional case: explicit method.

The methods introduced in (2·1) for the one dimensional system of conservation laws (2·1·1) will now be extended to the two dimensional system of conservation laws

$$(2\cdot4\cdot1) \qquad \frac{\partial u}{\partial t} + \frac{\partial f}{\partial x} + \frac{\partial g}{\partial y} = 0,$$

where u is once again the unknown column vector with n components, f and g are functions of the vector u, and so the system in non-conservation form is

$$(2\cdot4\cdot2) \qquad \frac{\partial u}{\partial t} + A\frac{\partial u}{\partial x} + B\frac{\partial u}{\partial y} = 0$$

where A and B are (n × n) matrices involving the unknown u. In [56], Richtmyer shows that the two step formulation of the Lax-Wendroff method for systems in two space dimensions is given by

$$(2\cdot4\cdot3) \quad \overset{*}{u}_{m+1} = \tfrac{1}{4}(\mu_x + \mu_y)u^m_{ij} - {}^p/4\,[H_x f^m_{ij} + H_y g^m_{ij}]$$

$$(2\cdot4\cdot4) \quad u_{m+1} = u_m - {}^p/2\,[H_x \overset{*}{f}{}^{m+1}_{ij} + H_y \overset{*}{g}{}^{m+1}_{ij}]$$

where

$$u_m = u^m_{ij} = u(ih, jh, mk)$$

and the operators $H_x$ and $H_y$ are defined by

$$H_x u^m_{ij} = u^m_{i+1j} - u^m_{i-1j}$$

and

$$H_y u^m_{ij} = u^m_{ij+1} - u^m_{ij-1}$$

Following the example of generalising the predictor-corrector scheme (in terms of 'a') in section (2·1), equations (2·4·3), (2·4·4) may be generalized similarly to give

$$(2\cdot4\cdot5)\quad \overset{*}{u}_{m+1} = \tfrac{1}{2}(\mu_x + \mu_y)u^m_{ij} - ap[H_x f^m_{ij} + H_y g^m_{ij}]$$

$$(2\cdot4\cdot6)\quad u_{m+1} = u_m - p[bH_x f^m_{ij} + cH_x \overset{*}{f}{}^{m+1}_{ij} + dH_y g^m_{ij} + eH_y \overset{*}{g}{}^{m+1}_{ij}],$$

where $b$, $c$, $d$ and $e$ are constants to be determined. Eliminating $\overset{*}{u}_{m+1}$ between (2·4·5) and (2·4·6), and expanding the resulting expression in terms of $u^m_{ij}$ and its derivatives and retaining terms up to, and including those of order $h^2$, it is found, for (2·4·5), (2·4·6) to be a second order correct approximation to (2·4·1), that

$$b = d = \tfrac{1}{2}\left(1 - \frac{1}{4a}\right)$$
$$c = e = \frac{1}{8a}$$

must be satisfied. The two dimensional analogue of (2·1·10) is, therefore

$$(2\cdot4\cdot7)\quad \begin{aligned} \overset{*}{u}_{m+1} &= \tfrac{1}{2}(\mu_x + \mu_y)u_m - ap[H_x f^m_{ij} + H_y g^m_{ij}] \\ u_{m+1} &= u_m - {}^{p}/_{2}\left[\left(1 - \frac{1}{4a}\right)(H_x f^m_{ij} + H_y g^m_{ij}) + \frac{1}{4a}(H_x \overset{*}{f}{}^{m+1}_{ij} + H_y \overset{*}{g}{}^{m+1}_{ij})\right]. \end{aligned}$$

A similar analysis will yield the second order method

$$\overset{*}{u}_{m+1} = \tfrac{1}{2}(\mu_x + \mu_y)u_m - ap\,[A_m H_x + B_m H_y]u^m_{ij}$$

$$u_{m+1} = u_m - {}^{p}/_{2}\left[\left(1 - \frac{1}{4a}\right)[A_m H_x + B_m H_y]u^m_{ij} + \frac{1}{4a}[\overset{*}{A}_{m+1} H_x + \overset{*}{B}_{m+1} H_y]\overset{*}{u}{}^{m+1}_{ij}\right]$$

where

$$A_m = A^m_{ij} = A(u^m_{ij}),$$

and

$$\overset{*}{A}_{m+1} = \overset{*}{A}{}^{m+1}_{ij} = A(\overset{*}{u}{}^{m+1}_{ij}),$$

and similarly for $B_m$ and $\overset{*}{B}_{m+1}$, as the two dimensional analogue of (2·1·11), for the system of equations (2·4·2).

## 2·5 Two dimensional case: The A.D.I. method.

Alternating Direction Implicit (A.D.I.) methods were first introduced by Peaceman and Rachford in 1955 [53] for the numerical solution of the heat conduction equation in two space variables. A.D.I. schemes have, in the past, been widely used for the solution of elliptic and parabolic partial differential equations. Their use in the field of hyperbolic systems of partial differential equations has only recently been considered, for example Kreiss [39], Gary [18], Gourlay and Mitchell [20], [21], [22]. The A.D.I. scheme introduced here is the two dimensional analogue of the one dimensional implicit scheme (2·2·6), (2·2·8) and the non-linear analogue of the A.D.I. method considered by Gourlay and Mitchell [22]. The method

$$(2\cdot5\cdot1)\quad [u_{m+1} + {}^{p}/_{4}\{H_x f_{m+1} + H_y g_{m+1}\} + {}^{p^2}/_{16}\, H_y \tilde{B}_{m+1} H_x f_{m+1}] =$$
$$= [u_m - {}^{p}/_{4}\,\{H_x f_m + H_y g_m\} + {}^{p^2}/_{16}\, H_y \tilde{B}_m H_x f_m]$$

where $\tilde{B}$ is defined by the relation

$$\tilde{B}(u).u = g(u)$$

is a second order accurate approximation to (2·4·1). By virtue of the non-linear terms $f_{m+1}$, $g_{m+1}$, this scheme (2·5·1) requires an iterative method to solve it at each point and is thus, in general, impracticable. We introduce, therefore, a method which does not suffer from the disadvantages of (2·5·1), namely a predictor-corrector method which employs a D'Jakonov [12] type factorization of (2·5·1). The resulting procedure is of the form

$$(2\cdot5\cdot2)\quad \overset{**}{u}_{m+1} = \tfrac{1}{2}(\mu_x + \mu_y)u_m - ap[H_x f^m_{ij} + H_y g^m_{ij}]$$

$$(2\cdot5\cdot3)\quad \overset{*}{u}_{m+1} = u_m - p[bH_y \overset{\sim**}{B}_{m+1}\overset{*}{u}_{m+1} + c(H_x f^m_{ij} + H_y g^m_{ij}) + dpH_y\tilde{B}H_x f^m_{ij}]$$

$$(2\cdot5\cdot4)\quad u_{m+1} = \overset{*}{u}_{m+1} - qpH_x\overset{\sim**}{A}_{m+1}\, u_{m+1}$$

where a, b, c, d and q are constants to be determined. Eliminating $\overset{*}{u}_{m+1}$, $\overset{**}{u}_{m+1}$ from (2·5·2), (2·5·3) and (2·5·4) and expanding the resulting expression using Taylor's series and comparing the coefficients with the expansion by Taylor's series of $u^{m+1}_{ij}$ in terms of $u^m_{ij}$ and its derivatives, the constants are found to satisfy the equations

$$b + c = \tfrac{1}{2}$$

$$c + q = \tfrac{1}{2}$$

$$bq + d = 0$$

$$2b = \tfrac{1}{2}$$

$$2q = \tfrac{1}{2}$$

$$4ba = \tfrac{1}{2}$$

$$4qa = \tfrac{1}{2}$$

if (2·5·2, 2·5·3, 2·5·4) is to the second order accurate. These equations have the solution $b = q = c = {}^1/_4$, $a = \tfrac{1}{2}$, $d = -{}^1/_{16}$. Hence the A.D.I. scheme is

$$(2\cdot5\cdot5)\quad \overset{**}{u}_{m+1} = \tfrac{1}{2}(\mu_x + \mu_y)u_m - {}^p/_2\,[H_x f^m_{ij} + H_y g^m_{ij}]$$

$$(2\cdot5\cdot6)\quad [I + {}^p/_4\, H_y\, \overset{\sim**}{B}_{m+1}]\overset{*}{u}_{m+1} = Q$$

$$(2\cdot5\cdot7)\quad [I + {}^p/_4\, H_x\, \overset{\sim**}{A}_{m+1}]u_{m+1} = \overset{*}{u}_{m+1},$$

where Q is the term on the right hand side of (2·5·1). The value of $u_{m+1}$ may be obtained from equations (2•5•5, 2•5•6, 2•5•7) by block tridiagonal matrix inversion techniques which may be obtained from The literature, for example Varga [68]. Equation (2•5•5) plays the role of the predictor formula yielding a first order approximation $\overset{**}{u}_{m+1}$ to $u(ih, jh, (m+1)k)$. Equation (2•5•6) when taken with (2•5•7) forms the alternating direction implicit part of the method. An A.D.I. method for the system of equations (2•4•2) may be derived in a similar manner. Its form is

$$\overset{**}{u}_{m+1} = \tfrac{1}{2}(\mu_x + \mu_y)u_m - {}^{P}/2[A_m H_x + B_m H_y]u_m$$

$$[I + {}^{P}/4\overset{**}{B}_{m+1}H_y]\overset{*}{u}_{m+1} = [I - {}^{P}/4[A_m H_x + B_m H_y] + {}^{P^2}/16B_m H_y A_m H_x]u_m$$

$$[I + {}^{P}/4\ \overset{**}{A}_{m+1}H_x]u_{m+1} = \overset{*}{u}_{m+1}$$

and is, once again, second order correct.

## 2·6 The "Explicit" Alternating Direction Method,

In [23] Gourlay and Mitchell considered several alternating direction methods for the solution of the equations

$$(2\cdot6\cdot1) \quad \frac{\partial u}{\partial t} = A(x,y,t)\frac{\partial u}{\partial x} + B(x,y,t)\frac{\partial u}{\partial y}$$

and briefly considered the extensions of these methods to the equations

$$\frac{\partial u}{\partial t} = \frac{\partial f(u)}{\partial x} + \frac{\partial g(u)}{\partial y}$$

and

$$\frac{\partial u}{\partial t} = A(u)\frac{\partial u}{\partial x} + B(u)\frac{\partial u}{\partial y}$$

(Note that the coefficient matrices $A$ and $B$ in $(2\cdot6\cdot1)$ do not depend upon $u$. We will have cause to consider the case when they do later in Chapter 4). We will consider modifications of these methods. The object of the schemes in [23] was to reduce the amount of computation and storage required by the methods by eliminating the nodal value $(ih, jh)$, thus reducing the nine point operators to four point operators at each time level. This scheme may be written for the system of conservation laws stated above as

$$\overset{**}{u}_{m+1} = \tfrac{1}{2}(\mu_x + \mu_y)u_m - {}^{p}/_{2}[H_x f_m + H_y g_m]$$

$$[\mu_y + {}^{p}/_{4} H_y \overset{\sim **}{B}_{m+1}]\overset{*}{u}_{m+1} = [\mu_y - {}^{p}/_{4} H_y \tilde{B}_m][\mu_x - {}^{p}/_{4} H_x \tilde{A}_m]u_m$$

$$[\mu_x + {}^{p}/_{4} H_x \overset{\sim **}{A}_{m+1}]u_{m+1} = \overset{*}{u}_{m+1}.$$

We point out, also, a difficulty which arises in the methods of [23] for the non-linear system of equations. The methods outlined there operate on a grid with mesh size 2h, <u>but</u> the coefficient matrices require to be calculated at mesh length h. The methods outlined in

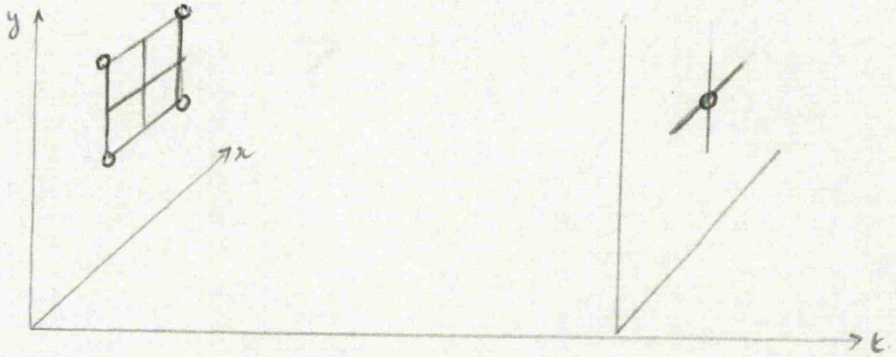
[23] could (in a complicated way) operate on a grid with mesh length h but the claimed "explicit" nature would then be lost on alternate points. We propose to rectify this position by introducing a scheme which works strictly on a 2h grid and which is strictly explicit and alternating in nature.

The above comments naturally apply to the predictor formula. It is therefore necessary to introduce the new predictor scheme

(2·6·2) $\overset{**}{u}_{m+1} = \mu_x\mu_y u^m_{ij} - P/2[\mu_y A(\mu_x u^m_{ij}) H_x u^m_{ij} + \mu_x B(\mu_y u^m_{ij}) H_y u^m_{ij}]$

where $\mu_x$, $\mu_y$ are the usual average difference operators, and where we are considering the system of equations in non-conservation form, $A(\mu_x u^m_{ij})$ means that the argument (in u) of A is evaluated as an average of $u_{i+1j}$ and $u_{i-1j}$.

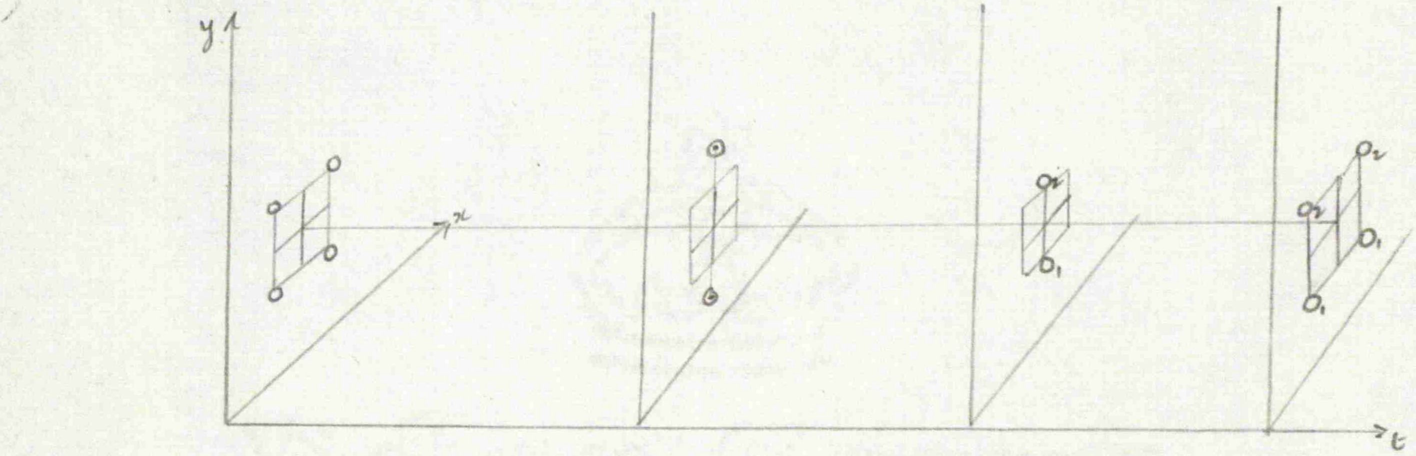
Equation (2·6·2) is evaluated for $i = 1, 3, \ldots N-1$ and $j = 1, 3, \ldots N-1$ (assuming N is even) and hence the method, diagramatically has the form



The split corrector scheme is then

(2·6·3) $[\mu_y + P/4 \overset{**}{B}_{m+1} H_y]\overset{*}{u}^{m+1}_{ij} = [\mu_y - P/4 B(\mu_x\mu_y u^m_{ij}) H_y][\mu_x - P/4 A(\mu_x u^m_{ij}] u^m_{ij}$

(2·6·4) $[\mu_x + P/4 \overset{**}{A}_{m+1}(\mu_y u^m_{ij}) H_x] u^{m+1}_{ij} = \overset{*}{u}^{m+1}_{ij}$, where care must be taken over the indices $ij$ in (2·6·3) and (2·6·4). For (2·6·3) $i = 1, 3, \ldots N-1$, $j = 1, 3, \ldots N-1$ whilst for (2·6·4) $i = 1, 3, \ldots N-1$ but $j = 2, 4, \ldots N-2$, so that the whole method (2·6·2, 2·6·3, 2·6·4) is represented diagramatically as



At the $(m+1)^*$ level $\theta_1$ and $\theta_2$ are determined in terms of nodal values $\theta$ at the m-level and the matrix $B^{**}$ is evaluated at the middle $\theta$ at the $^{**}$ level. At the $(m+1)$ level the two $\theta_1$'s are determined in terms of the $\theta_1$ at the $(m+1)^*$ level and the matrix $A^{**}$ is evaluated with its arguments as an average of the middle and lower $\theta$'s at the $^{**}$ level, whilst the two $\theta_2$'s at the $(m+1)$ level are evaluated in terms of $\theta_2$ at the $(m+1)^*$ level, and the matrix $A^{**}$ is evaluated with its argument as an average of the middle and upper $\theta$'s at the $^{**}$ level. As can be seen there is a mesh size 2h operating, and the values at the $(m+1)$ level are explicitly obtained. (The solution at the $m+1$ level proceeds with $j = 2$ and $i = 1, 3, \ldots N-1$, $j = 4$, $i = 1, 3, \ldots N-1$, $\ldots$ $j = N-2$, $i = 1, 3, \ldots N-1$. The problem of boundary points does not arise as points on the boundary are not required, since the $(m+1)$ level formula is not used for $j = 0$ and $j = N$ as theoretical data is available at the $(m+1)$ level.

A similar scheme may be derived for the system of conservation laws

$$\frac{\partial u}{\partial t} + \frac{\partial f}{\partial x} + \frac{\partial g}{\partial y} = 0,$$

where once again care must be taken in the evaluation of the matrices $\tilde{A}$ and $\tilde{B}$. We find that the scheme is of the form

$$(2{\cdot}6{\cdot}5)\quad \overset{**}{u}_{m+1} = \mu_x\mu_y u^m_{ij} - {}^p/_2[\mu_y H_x f^m_{ij} + \mu_x H_y g^m_{ij}]$$

$$(2{\cdot}6{\cdot}6)\quad [\mu_y + {}^p/_4 H_y \tilde{B}(\mu_x u^{**m+1}_{ij})]u^{*m+1}_{ij} = [\mu_y - {}^p/_4 H_y \tilde{B}(\mu_x u^m_{ij})][\mu_x - {}^p/_4 H_x \tilde{A}(u^m_{ij})]u^m_{i}$$

$$(2{\cdot}6{\cdot}7)\quad [\mu_x + {}^p/_4 H_x \tilde{A}(\mu_x\mu_y u^{**m+1}_{ij})]u^{m+1}_{ij} = u^{*m+1}_{ij}$$

Comments applied to (2·6·2, 2·6·3, 2·6·4) apply to (2·6·5, 2·6·6, 2·6·7). We note the slightly more complicated averaging of the matrices here by virtue of the fact that the matrices are always preceded by a difference operator. The method (2·6·5, 2·6·6, 2·6·7) is again explicit and operates on a grid of mesh size 2h. We have not given a full derivation of the above schemes but merely stated the result of the derivation, which proceeded along similar lines to the derivation of the Alternating Direction Implicit methods of section (2·5). The predictor level is used only for the evaluation of the arguments of the matrices. When the argument involves a point outside the region of computation, a boundary replacement must be used, which will be described in the next chapter. This applies in particular to $\tilde{A}$ in (2·6·7) and $\tilde{B}$ in (2·6·6).

## 2·7 Stability of schemes of sections 2·4, 2·5 and 2·6.

Following Richtmyer [58] we consider the stability of the above schemes for the system of equations in fluid dynamics in Non-conservation Eulerian form

$$(2{\cdot}7{\cdot}1) \quad \frac{\partial\phi}{\partial t} + A(\phi)\frac{\partial\phi}{\partial x} + B(\phi)\frac{\partial\phi}{\partial y} = 0,$$

where

$$\phi = \begin{bmatrix} \rho \\ u \\ v \\ p \end{bmatrix}, A(\phi) = \begin{bmatrix} u & \rho & \cdot & \cdot \\ \cdot & u & \cdot & 1/\rho \\ \cdot & \cdot & u & \cdot \\ \cdot & \rho c^2 & \cdot & u \end{bmatrix}, B(\phi) = \begin{bmatrix} v & \cdot & \rho & \cdot \\ \cdot & v & \cdot & \cdot \\ \cdot & \cdot & v & 1/\rho \\ \cdot & \cdot & \rho c^2 & v \end{bmatrix}$$

Here $\rho, u, v$ and $p$ represent the density, x-component of velocity, y-component of velocity and pressure of the fluid respectively and $c$ is the local sound speed in the fluid. Considering (2·4·4) for the solution of this system (2·7·1), it is found on eliminating $\overset{*}{u}_{m+1}$ from the two equations of (2·4·4) and linearising the result that

$$u_{m+1} = u_m - {}^{p}/2[AH_x + BH_y][(1 - \tfrac{1}{4a})u_m + \tfrac{1}{8a}(\mu_x + \mu_y)u_m] +$$
$$\frac{p^2}{8}[AH_x + BH_y]^2 u_m,$$

where $A$ and $B$ are assumed to be locally constant. After the usual Fourier transform of the variables, the amplification matrix $G$ is given by

$$G = I - \sqrt{-1}p\ [A\sin\alpha h + B\sin\beta h][(1 - \tfrac{1}{4a}) + \tfrac{1}{8a}(\cos\alpha h + \cos\beta h)] -$$
$$\frac{p^2}{2}[A\sin\alpha h + B\sin\beta h]^2$$

where $\alpha$ and $\beta$ are real numbers. Let $Z$ denote the matrix $A\sin\alpha h + B\sin\beta h$. Then if $\lambda_j$ is an eigenvalue of $Z$ the eigenvalues of $G^*G$, where $G^*$ is the Hermitian transpose of $G$, are given by

$$g = \left\{1 - \frac{p^2\lambda_3^2}{2}\right\}^2 + p^2\lambda_3^2 \left[\left(1-\frac{1}{4a}\right) + \frac{1}{8a}(\cos\alpha h + \cos\beta h)\right]^2 .$$

For the Van Neumann condition for stability to be satisfied we require $|g| \leqslant 1$ and therefore

$$\frac{p^2\lambda_3^2}{4} \leqslant 1 - \left[\left(1 - \frac{1}{4a}\right) + \frac{1}{8a}(\cos\alpha h + \cos\beta h)\right]^2$$

for all real $\alpha$ and $\beta$. That is

$$p^2 \leqslant \min_{\alpha,\beta} \left\{ \frac{1}{\lambda_3^2}\left[4 - \frac{1}{16a^2}\left[8a - 2 + (\cos\alpha h + \cos\beta h)\right]^2\right]\right\} .$$

Because $\lambda_3$ is dependent on $\alpha, \beta, A$ and $B$ we consider the stability for a particular physical problem, namely (2·7·1). For the matrices $A(\phi)$ and $B(\phi)$ of (2·7·1) it may readily be shown that the maximum eigenvalue of the matrix $Z^2$ is given by

$$(|q| + c)^2(\sin^2\beta h + \sin^2\alpha h),$$

where

$$|q| = |(u^2 + v^2)^{\frac{1}{2}}| .$$

In this case, therefore, we have the stability condition for (2·4·4)

$$p^2(|q| + c)^2 \leqslant \min_{\alpha,\beta} \left\{ \frac{4 - \frac{1}{16a^2}\left[8a - 2 + (\cos\alpha h + \cos\beta h)\right]^2}{\sin^2\alpha h + \sin^2\beta h} \right\}$$

$$= M(a).$$

By theoretical arguments, or by a simple scanning run on a computer, it may be shown that $M(a)$ is negative for $a < \frac{1}{4}$, has a maximum at $a = \frac{1}{4}$, and decreases steadily in the range $\frac{1}{4} \leqslant a \leqslant \infty$. A graph showing the behaviour of $M(a)$ is given in figure (2·7·1).

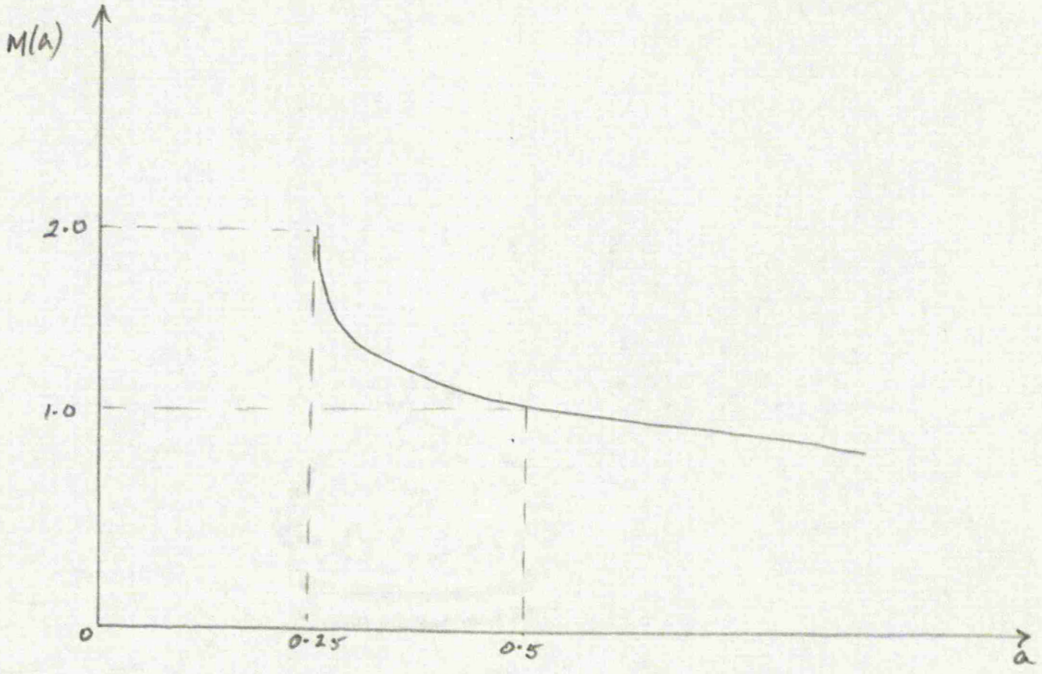


figure (2.7.1)

The stability condition for the A.D.I. method gives unconditional stability, since the linearised analysis reduces to considering the stability of the A.D.I. method of Gourlay and Mitchell [22]. Similarly the stability analysis of the Explicit Alternating Direction method reduces to the stability analysis considered by Gourlay and Mitchell [23]. This is so because the present method differs from the method considered in [23] only in the introduction of a predictor level, and the latter is used <u>only</u> in the evaluation of the arguments of the matrices. Since these are assumed to be constant, the new predictor formula does not affect the stability characteristics and hence the Explicit Alternating Direction methods are unconditionally stable.

## 2·8 Conservation properties of the general predictor-corrector scheme.

Finite difference methods are said to be in conservation difference form if they can be written as

$$u_{m+1} = u_m + D_x^h F,$$

where $D_x^h$ is a central difference operator and $F$, which is a function of $u_m$ and possibly difference operators, has the property that if all the arguments are put equal, it reduces to the function $f$ in the differential equation. Lax [42] has shown that the solution of a consistent difference equation written in conservation form is a weak solution of the conservation law

$$(2\cdot8\cdot1) \quad \frac{\partial u}{\partial t} + \frac{\partial f}{\partial x} = 0.$$

From this we see that the finite difference analogue of the quantity $\int u dx$ is directly related to the difference analogue of the integrated flux $\int f(u)dt$; that is if we integrate (2·8·1) with respect to $t$ and $x$ between $t_1$ and $t_2$ and $x_1$ and $x_2$ to give

$$\int_{x_1}^{x_2} u dx \Big|_{t_1}^{t_2} + \int_{t_1}^{t_2} f(u) dt \Big|_{x_1}^{x_2} = 0,$$

then the difference analogue of this expression, namely

$$h \sum_j u_j \Big|_{t_1}^{t_2} + k \sum_j f_j = 0$$

is exactly satisfied since summation over all net points between $x_1$ and $x_2$ yields a telescoping sum of the quantities $f_j$. Gary [18] on the other hand shows that the Lax-Wendroff scheme when written as a one step method [45] is conservative by considering the solution of the equations of hydrodynamics, viz.

$$(2\cdot8\cdot2) \quad \frac{\partial w}{\partial t} + \frac{\partial f(w)}{\partial x} = 0$$

where

$$w = \begin{bmatrix} \rho \\ e \\ m \end{bmatrix}, \quad \text{and} \quad f(w) = \begin{bmatrix} m \\ \frac{yem}{\rho} - (y-1)\frac{m^3}{\rho^2} \\ (y-1)e - (y-3)\frac{m^2}{2\rho} \end{bmatrix},$$

with $\rho, e, m$ the mass, energy and momentum, respectively, per unit volume of the gas and $y$ the ratio of the specific heats of the gas. The method is conservative if the total mass, energy and momentum is the same at two time levels, except for the "flux" throughout the boundary.

On applying the Lax-Wendroff method to (2·9·2) and summing over the domain (of computation), we obtain

$$h \; \ell \sum_{j=1}^{N} w_j^{n+1} = h \sum_{j=1}^{N} w_j^n + \frac{k}{2}\left[f_1^n + f_2^n - f_{N-1}^n - f_N^n\right] + \frac{k^2}{2h}\left[q_N - q_1\right],$$

where

$$q_j = \left[\frac{A(w_{j+1}^n) + A(w_j^n)}{2}\right]\left[f(w_{j+1}^n - f(w_j^n)\right].$$

and $A$ is the Jacobian matrix of $f$ with respect to the components of $w$, namely

$$A = \begin{bmatrix} u & \cdot & \rho \\ \cdot & u & yp \\ \cdot & 1/\rho & u \end{bmatrix}.$$

We will now show that the general predictor-corrector method still preserves the property of conservation. Consider the general predictor-corrector scheme applied to (2·8·2), viz.

$$(2\cdot8\cdot3) \quad \overset{*}{w}_{m+1} = \mu_x w_i^m - apH_x f_i^m,$$

$$(2\cdot8\cdot4) \quad w_{m+1} = w_m - {}^p/2\left[\left(1 - \frac{1}{4a}\right)H_x f_i^m + \frac{1}{4a} H_x f_i^{*m+1}\right].$$

Multiplying (2·8·4) throughout by h and expanding the difference operator in terms of differences, we obtain on summing over the range,

$$(2\cdot8\cdot5) \quad h \sum_{i=1}^{N-1} w_i^{m+1} = h \sum_{i=1}^{N-1} w_i^m - \frac{k}{2} \sum_{i=1}^{N-1}\left[\left(1 - \frac{1}{4a}\right)(f_{i+1}^m - f_{i-1}^m) + \right.$$

$$\left.\frac{1}{4a}\left(f_{i+1}^{*m+1} - f_{i-1}^{*m+1}\right)\right].$$

The second sum on the right hand side of (2·9·5) gives us

$$\frac{k}{2}\left[\left(1-\frac{1}{4a}\right)(f_0^m + f_1^m - f_{N-1}^m - f_N^m) + \frac{1}{4a}(f_0^{*m+1} + f_1^{*m+1} - f_{N-1}^{*m+1} - f_N^{*m+1})\right],$$

which is the "flux" through the boundaries at the m'th and (m+1)* levels. We can see from this by varying 'a' that the flux is a weighted average over different levels and moreover, a = ½ gives a "mean" flux over two levels. The natural question to ask is "Is it possible to choose 'a' such that this flux is zero?" A simple analysis will quickly yield a negative answer to this question. Thus (2·8·5) states that the sum of mass, energy and momentum is the same at two time levels except for the "flux" through the boundaries; that is, the general predictor-corrector scheme is conservative.

## 2·9 Numerical Results.

### 1. One dimensional case

In order to compare the methods derived from the general predictor-corrector method (that is, for particular values of a), the simple problem

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}(\tfrac{1}{2}u^2) = 0$$

subject to the initial condition

$$u(x,0) = x, \quad 0 \leqslant x \leqslant 1,$$

which has the solution

$$(2\cdot9\cdot10) \quad u(x,t) = \frac{x}{1+t}$$

was considered. The interval under consideration ([0,1]) was divided into steps of $h = 0\cdot1$. The methods were run on a computer for $p = 0\cdot3$, $0\cdot6$ and $1\cdot0$ and the answers given at 100 and 300 time steps respectively. The errors at $x = \frac{1}{2}$, the difference between theoretical and computed solutions, are quoted in table (2·9·1). A comparison was made between the explicit and implicit methods.

### Boundary data.

In order to solve the problem the finite difference methods require boundary data on $x = 0$ and $x = 1$. This data was provided by substituting theoretical values of $u$ obtained from (2·9·1) at $x = 0$ and $x = 1$.

### 2. Two dimensional case

To compare the methods in the two space dimensional case, the problem

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{1}{4}u^2\right) + \frac{\partial}{\partial y}\left(\frac{1}{4}u^2\right) = 0,$$

subject to the initial condition

$$u(x,y,0) = \tfrac{1}{4}(x+y)^2$$

was chosen. This problem has the theoretical solution

$$u(x,y,t) = \left\{\frac{1-\sqrt{1+(x+y)t}}{t}\right\}^2 .$$

The region $0 \leq x, y \leq 1$ was considered and a square grid with mesh spacing $h = 0{\cdot}1$ was superimposed on the region. The two dimensional explicit and A.D.I. methods were run for varying $p$ and the errors are quoted at the point $(\tfrac{1}{2},\tfrac{1}{2})$ in table (2·9·2).

Boundary data.

Once again, boundary values are required in order to solve the finite difference schemes. In the case of the explicit method, data is required on the planes $x = 0$, $y = 0$, $x = 1$, $y=1$, for $0 < t \leq T$ (say). This data was again obtained by using theoretical values of $u(x,y,t)$ at the appropriate boundary points. In the case of the A.D.I. method, theoretical data was incorporated at the predictor boundary level and the $(m+1)^{\text{th}}$ level. The data required for the intermediate level ( the * level of the A.D.I. scheme) was provided by using (2·5·7) at $j = 0$ and $N$, for $i = 1, 2, \ldots N-1$, in order to obtain

$$u_{io}^{m+1^*} \text{ and } u_{iN}^{m+1^*}.$$

| $p$ | Number of time steps | Lax-Wendroff $a = \frac{1}{4}$ | $a = \frac{1}{2}$ | $a = 1$ | Implicit |
|---|---|---|---|---|---|
| $0\cdot 3$ | 100 | $\cdot 62 \times 10^{-4}$ | $\cdot 60 \times 10^{-4}$ | $\cdot 92 \times 10^{-3}$ | $\cdot 14 \times 10^{-3}$ |
| | 300 | $\cdot 94 \times 10^{-4}$ | $\cdot 51 \times 10^{-4}$ | $\cdot 76 \times 10^{-3}$ | $\cdot 19 \times 10^{-4}$ |
| $0\cdot 6$ | 100 | $\cdot 15 \times 10^{-4}$ | $\cdot 17 \times 10^{-3}$ | $\cdot 37 \times 10^{-2}$ | $\cdot 10 \times 10^{-2}$ |
| | 300 | $\cdot 11 \times 10^{-2}$ | $\cdot 85 \times 10^{-3}$ | $\cdot 11 \times 10^{-3}$ | $\cdot 47 \times 10^{-3}$ |
| $1\cdot 0$ | 100 | $\cdot 11 \times 10^{-2}$ | $\cdot 11 \times 10^{-2}$ | $\cdot 57 \times 10^{-3}$ | $\cdot 16 \times 10^{-2}$ |
| | 300 | $\cdot 36 \times 10^{-3}$ | $\cdot 22 \times 10^{-2}$ | $\cdot 25 \times 10^{-2}$ | $\cdot 29 \times 10^{-3}$ |

table (2·9·1)

| $p$ | Number of Time steps | Lax-Wendroff $a = \frac{1}{4}$ | $a = \frac{1}{2}$ | A.D.I. method |
|---|---|---|---|---|
| $0\cdot 3$ | 50 | $\cdot 78 \times 10^{-3}$ | $\cdot 60 \times 10^{-3}$ | $\cdot 49 \times 10^{-3}$ |
| | 100 | $\cdot 80 \times 10^{-3}$ | $\cdot 59 \times 10^{-3}$ | $\cdot 49 \times 10^{-3}$ |
| $0\cdot 6$ | 50 | $\cdot 76 \times 10^{-3}$ | $\cdot 59 \times 10^{-3}$ | $\cdot 47 \times 10^{-3}$ |
| | 100 | $\cdot 72 \times 10^{-3}$ | $\cdot 52 \times 10^{-3}$ | $\cdot 44 \times 10^{-3}$ |
| $1\cdot 0$ | 50 | $\cdot 59 \times 10^{-3}$ | $\cdot 53 \times 10^{-3}$ | $\cdot 44 \times 10^{-3}$ |
| | 100 | $\cdot 68 \times 10^{-3}$ | $\cdot 45 \times 10^{-3}$ | $\cdot 48 \times 10^{-3}$ |

table (2·9·2)

## 2·10 Comments on the methods introduced in Chapter 2

The parameter 'a' considered in the derivation of the explicit predictor-corrector scheme plays a significant part in the method. From the stability analysis of (2·3) it was seen that 'a' had to be greater than or equal to $\frac{1}{4}$ and also $p|\lambda| \leqslant \frac{1}{\sqrt{a}}$ where $\lambda$ was the maximum modulus eigenvalue of the Jacobian matrix A.

The question therefore arises 'what is the significance of 'a'?' If the predictor of (2·1·10) is expanded using Taylor's series it is found that

$$\overset{*}{u}_{m+1} = u_m - 2ak\frac{\partial f}{\partial x} + O(h^2),$$

which is seen to be a first order approximation to u located at the position $(m + 2a)k$; that is, the position of the predictor level is $(m + 2a)k$ in time. Two points therefore arise,

1. It is forbidden by the stability analysis to have the prediction within the interval mk and $(m + \frac{1}{2})k$.

2. Prediction may be made ahead of the $(m + 1)k$ level. For example $a = 1$ implies that the prediction is made at $(m + 2)k$, when the predictor-corrector scheme reduces to

$$\overset{*}{u}_{m+1} = \mu_x \overset{*}{u}_m - pH_x f_i^m$$

$$u_{m+1} = u_m - {}^p/_8 H_x[3f_i^m + \overset{*}{f}{}_i^{m+1}],$$

for the dimensional system of conservation laws.

The explicit schemes of (2·1) bear a strong resemblance to several finite difference methods of solving the ordinary differential equation $\frac{dy}{dx} = f(x,y)$. For example, the Lax-Wendroff method (2·1·2, 2·1·3) is similar to the "improved Euler method",

$$y_{n+\frac{1}{2}} = y_n + \frac{h}{2}f_n,$$

$$y_{n+1} = y_n + hf_{n+\frac{1}{2}},$$

where $f_n = f(x_n, y_n)$, whereas the scheme (2·1·10) with $a = \frac{1}{2}$ mirrors the "improved Euler-Cauchy" method,

$$\overset{*}{y}_{n+1} = y_n + hf_n,$$

$$y_{n+1} = y_n + \tfrac{1}{2}h[f_n + \overset{*}{f}_{n+1}].$$

An important point concerning boundary values arises as a result of the computation. The finite difference methods introduced are of higher order in the space dimensions than the original differential system, thereby creating the situation where the difference system requires more boundary data than the differential system, in other words, the difference system is underdetermined. We know, in theory, that the addition of extra boundary values at $x = 1$ when the Jacobian matrix has positive eigenvalues, automatically over-determines the differential problem which already has initial data on $t = 0$ and boundary data at $x = 0$. It can be seen from the results (although not quoted in tables (2·9·1), (2·9·2)), that the errors nearer the boundary are significantly larger than those nearer the centre of the field of computation. The question arises, therefore, 'can we introduce either some method by which one may eliminate, or "smooth", the increased errors near the boundary, or indeed, devise some technique which dispenses with the need for an upper boundary?' As the experiments of Richtmyer and Morton [59], Gary [18] and Parter [52] have shown, when a numerical method is implemented, considerable care has to be taken when inserting extra data. For example, the paper by Richtmyer and Morton [59] considers several extrapolation schemes by which the unknown boundary point may be determined in the explicit methods.

In the following chapter, we will investigate this problem of boundary data and introduce methods by which more accurate results may be obtained, without overdetermining the problem.

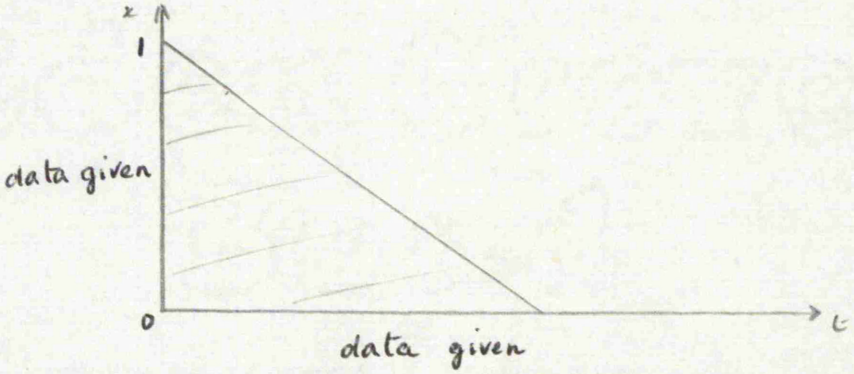
# CHAPTER III

# BOUNDARY TECHNIQUES

Gourlay and Morris [70]

## 3.1 Introduction.

It was found in the previous chapter that the data given for an initial value problem namely initial data and data along a line for increasing time (for example the line $x = 0$) was usually not sufficient to determine the solution of the finite difference schemes. To surmount this difficulty we introduced theoretical data on $x = 1$, but, in so doing, overdetermined the problem. The reason for the need of data at the second boundary in explicit methods is obvious:- not introducing such data would yield the scheme capable of deriving a solution in a wedge shaped region only, -



Similarly the reason for introducing 'extra' data in the implicit methods is obvious, - not to do so would yield 'too many unknowns and too few equations to solve' at each time level.

In the numerical examples of section (2·10), we noted that the errors near the 'upper' boundary (that is, $x = 1$) were significantly larger than those in the interior of the region. We surmised that these errors arose out of the method of inserting the extra boundary data. There are two lines of approach open to us in order, either to reduce this error or to eliminate it altogether (that is, bring the magnitude of the error at the boundary into line with that of the interior). An indiciation of the first line of approach is given by the use of predictor-corrector methods in the solution of ordinary differential equations which has already been mentioned in (2·2). This amounts to iterating the corrector formula several times at the same time location in order to smooth out the boundary errors over the whole region. Such a scheme was considered by Gary [18] for the solution of equations of the form $\frac{\partial w}{\partial t} + A(w)\frac{\partial w}{\partial x} = 0$. The second line of approach is to endeavour to avoid introdusing larger errors at the boundary by using a method similar to that considered by Morton and Richtmyer. These schemes are extrapolation procedures by which the unknown boundary data is given as a linear combination of values inside the region of computation.

In section (3·7·8) an important application of this technique will be investigated in conjunction with a classical procedure known as "deferred approach to the limit" or "Richardson's Extrapolation".

## 3·1 Iteration of the Corrector.

Consider the one dimensional system of conservation laws,

$$(3\cdot1\cdot1)\quad \frac{\partial u}{\partial t} + \frac{\partial f}{\partial x} = 0,$$

where $u$ is an unknown vector function and $f$ is a known vector function of the components of $u$, together with the explicit general predictor-corrector scheme, (2·1·10), namely,

$$(3\cdot1\cdot2)\quad \overset{*}{u}_{m+1} = \mu_x u_m - ap H_x f_i^m$$

$$(3\cdot1\cdot3)\quad u_{m+1} = u_m - {}^{p}/2\left[\left(1 - \frac{1}{4a}\right)H_x f_i^m + \frac{1}{4a}H_x \overset{*}{f}{}_i^{m+1}\right]$$

where the notation

$$u_m = u_i^m = u(ih, mk) \quad i = 0, 1, \dots N, \ m = 0, 1, \dots$$

has once again been used and $H_x u_i^m = u_{i+1}^m - u_{i-1}^m$ is again the central space difference operator. The mesh ratio is $p = {}^{k}/h$ where $k$ and $h$ are the mesh spacings in the time and space directions respectively.

Formulae (3·1·2, 3·1·3) may be regarded as generating the predictor-corrector scheme

$$\overset{*}{u}_{m+1} = \mu_x u_m - ap H_x f_i^m,$$

$$u_{m+1}^{(j+1)} = u_m - {}^{p}/2\left[\left(1 - \frac{1}{4a}\right)H_x f_m + \frac{1}{4a}H_x f_{m+1}^{(j)}\right]$$

for $j = 1, 2 \dots$ where

$$f_{m+1}^{(1)} = \overset{*}{f}_{m+1}$$

and the sequence of iterates $\{u_{m+1}^{(j+1)} \quad j = 1, 2, \dots\}$ is required to tend to the solution of (3·1·1) at the nodes $(ih, (m+1)k)$. We see therefore that we are really finding the solution of

(3·1·5) $u_{n+1} = u_n - {}^{p}/2\left[\left(1 - \frac{1}{4a}\right)H_x f_n + \frac{1}{4a} H_x f_{n+1}\right]$

by an iterative procedure which we note is similar to the implicit scheme (2·2·3, 2·2·4). If we now expand the right hand side of (3·1·5) in terms of $u_n$ and its derivatives with respect to $x$ and $t$ using Taylor's series and considering terms up to and including those of $O(k^2)$ in the resulting expression, the expansion

$$u_{n+1} = \left[u - k\frac{\partial f}{\partial x} - \frac{k^2}{4a}\frac{\partial}{\partial x}\left(\frac{\partial f}{\partial t}\right)\right]_n + O(k^3)$$

is obtained. If we use (3·1·1), we obtain

$$u_{n+1} = \left[u + k\frac{\partial u}{\partial t} + \frac{k^2}{4a}\ \frac{\partial^2 u}{\partial t^2}\right]_n + O(k^3).$$

Whilst every replacement is a consistent one, only the choice $a = \frac{1}{2}$ gives a corrector which may converge to the order $(O(k^2))$ of the overall scheme.

We will now demonstrate the convergence of the iterated corrector in (3·1·4). If we subtract (3·1·5) from the corrector in (3·1·4), the equation

(3·1·6) $u_{n+1}^{(j+1)} - u_{n+1} = -{}^{p}/8a\, H_x[f_{n+1}^{(j)} - f_{n+1}]$

is obtained. In order to carry the analysis further, we are forced to linearize this equation, that is, we assume the Jacobian matrix $A$ of $f$ to be locally constant at any time level. That is, (3·1·6) may be written as

(3·1·7) $u_{n+1}^{(j+1)} - u_{n+1} = -{}^{Ap}/8a\, H_x[u_{n+1}^{(j)} - u_{n+1}]$

Using the notation

$$e_i^{j+1} = u_{n+1}^{(j+1)} - u_{n+1},$$

and letting $E_{j+1}$ denote the vector $(e_1^{j+1}, \dots e_N^{j+1})^T$, then considering the totality of equations of the form (3·1·6) for consecutive iterates and assuming periodic boundary conditions, we may write the resulting system of equations in the matrix form, -

(3·1·8) $E_{j+1} = A_1 E_j$,

where $A_1$ is the block tridiagonal matrix

$$\begin{bmatrix} 0 & Ap/8a & & & \\ -Ap/8a & 0 & Ap/8a & & \\ & \ddots & \ddots & \ddots & \\ & & & & Ap/8a \\ & & & & 0 \\ & & & -Ap/8a & \end{bmatrix}$$

The iterative process will converge if

$$\rho(A_1) < 1$$

where $\rho(A_1)$ denotes the spectral radius of $A_1$. It is easily seen that if $\lambda$ is an eigenvalue of $A$ then the corresponding eigenvalue of $A_1$ will be

$$\lambda_1 = 2\sqrt{-1}\ {}^{\lambda p}/8a\ \cos\frac{s\pi}{N+1}, \quad \text{for } s = 1, \dots N$$

and the maximum modulus eigenvalue of $A_1$ corresponding to $\lambda$ is

$$\le \frac{|\lambda|p}{4a}$$

Therefore $\displaystyle\lim_{j \to \infty} E_j \to 0$ and so $\displaystyle\lim_{j\to\infty} e^j \to 0$ leading to

$$\lim_{j\to\infty} u_{m,t}^{(j)} \to u_{m,t}$$

if and only if

$$|\lambda_1| < 1 \text{ which implies } p|\lambda| < 4a.$$

Hence we require $p|\lambda| < 4a$ for convergence for all eigenvalues of $A$. This condition is satisfied by virtue of the stability requirements of (2·1·10). We see that the most satisfactory predictor-corrector process is obtained when $a = \frac{1}{2}$ for only in this case do the iterates converge to the order of the overall method. This scheme is given by

$$\begin{aligned} \overset{*}{u}_{m+1} &= \mu_x u_m - {}^{p}/_{2}\, H_x f_m \\ u_{m+1}^{(j+1)} &= u_m - {}^{p}/_{4}\, H_x[f_m + f_{m+1}^{(j)}] \, . \end{aligned} \tag{3·2·9}$$

Hence the use of this iterative procedure will, providing the conditions for convergence are satisfied, smooth out the errors introduced at the boundary by substituting theoretical data there. The scheme derived above for (3·2·1) may be extended in an obvious way to the solution of the equations written in non-conservation form, namely

$$\frac{\partial u}{\partial t} + A(u)\frac{\partial u}{\partial x} = 0.$$

For completeness we state the two dimensional analogues of the above methods. Their derivation follows similar lines to the above and hence will not be repeated. For the equations

$$\frac{\partial u}{\partial t} + \frac{\partial f}{\partial x} + \frac{\partial g}{\partial y} = 0$$

where $f$ and $g$ are known vector functions of the unknown vector $u$, the iterative predictor-corrector method is

$$\begin{aligned} \overset{*}{u}_{m+1} &= \tfrac{1}{2}(\mu_x + \mu_y)u_m - ap[H_x f_m + H_y g_m] \\ u_{m+1}^{(j+1)} &= u_m - {}^{p}/_{2}[(1 - \tfrac{1}{4a})(H_x f_m + H_y g_m) + \tfrac{1}{4a}(H_x f_m^{(j)} + H_y g_m^{(j)})], \\ j &= 1, 2, \ldots \end{aligned} \tag{3·2·10}$$

where $u_m = u_{ij}^m = u(ih, jh, mk)$, with $H_x$, $H_y$, $\mu_x$, $\mu_y$ the usual difference operators, in the $x$ and $y$ directions. The corresponding scheme for the two dimensional equations in non-conservation form

$$\frac{\partial u}{\partial t} + A\frac{\partial u}{\partial x} + B\frac{\partial u}{\partial y} = 0,$$

is

$$\begin{aligned} &\overset{*}{u}_{n+1} = \tfrac{1}{2}(\mu_x + \mu_y)u_n - ap[A_nH_x + B_nH_y]u_n \\ (3\cdot 1\cdot 11)\quad &u^{(j+1)}_{n+1} = u_n - {}^{p}/_{2}[(1 - \tfrac{1}{4a})\{A_nH_x + B_nH_y\}u_n + \tfrac{1}{4a}\{A^{(j)}_{n+1}H_x + B^{(j)}_{n+1}H_y\}u^{(j)}_{n+1}] \\ &j = 1, 2, \ldots \end{aligned}$$

Analyses for (3·1·10) and (3·1·11) similar to the one dimensional case again yield the "optimum" value of 'a' to be $\frac{1}{2}$ and the convergence condition

$$\max\,[p|\lambda_A|,\ p|\lambda_B|] \leq 1,$$

which is satisfied by virtue of the stability requirements of the predictor-corrector methods of (2·4).

## 3·3 Extrapolation Boundary Techniques

We now adopt the second line of approach suggested in the introduction of this chapter. That is, we try to produce a scheme which will introduce the data at the boundary without introducing errors which are inconsistent with those arising from the method used in the interior of the region. We therefore try to develop a boundary replacement which will give us a truncation error which is the same as that in the interior of the region.

Consider once again the general predictor-corrector method

$$(3\cdot3\cdot1) \quad \overset{*}{u}_{n+1} = \mu_x u_n - apH_x f_n$$

$$(3\cdot3\cdot2) \quad u_{n+1} = u_n - {}^{p}/2\left[\left(1 - \frac{1}{4a}\right)H_x f_n + \frac{1}{4a}H_x \overset{*}{f}_{n+1}\right] .$$

It will be convenient for the consideration of this section to write $H_x$ as $\Delta_x + \nabla_x$ and $\mu_x = \frac{1}{2}(\Delta_x - \nabla_x + 2)$ so that (3·3·1), (3·3·2) may be written as

$$(3\cdot3\cdot3) \quad \overset{*}{u}_{n+1} = \tfrac{1}{2}(\Delta_x - \nabla_x + 2)u_n - ap\,(\Delta_x + \nabla_x)f_n$$

$$(3\cdot3\cdot4) \quad u_{n+1} = u_n - {}^{p}/2\left[\left(1 - \frac{1}{4a}\right)(\Delta_x + \nabla_x)f_n + \frac{1}{4a}(\Delta_x + \nabla_x)\overset{*}{f}_{n+1}\right].$$

These equations are obviously valid for $i = 1, 2, .. N-1$ where the boundaries of the region are labelled $i = 0$ and $i = N$. If we consider the differential problem

$$\frac{\partial u}{\partial t} + \frac{\partial f}{\partial x} = 0,$$

with the Jacobian matrix $A$ having positive eigenvalues, then the differential problem provides data on the initial line and on the 'lower' boundary line $i = 0$. Hence on this line theoretical values of $\overset{*}{u}{}_0^{n+1}$ and $u_0^{n+1}$ may be introduced. The problem of the extra boundary data therefore applies to the 'upper' boundary $i = N$ and in what follows we will assume the problem to be of this form (that is has positive eigenvalues), and derive a procedure for finding the extra boundary data on $i = N$. This consideration is for convenience only and a similar method may be derived for the problem with negative eigenvalues when a boundary procedure, slightly different from the one we shall derive, for $i = 0$, is needed. To obtain $u_N^{*m+1}$ and $u_N^{m+1}$ by formulae (3·3·3) and (3·3·4), we would require a value of the function $f$ at $f_{N+1}^{m}$ and $f_{N+1}^{*m+1}$ respectively, values which are not available to us. However, if we use the relations between the difference operators, as given below, to substitute for $\Delta_x$ in terms of $\nabla_x$ in (3·3·3) and (3·3·4) for $i = N$ only, we will obtain a procedure which will yield $u_N^{*m+1}$ and $u_N^{m+1}$ in terms of values of $u_i^m$ and $u_i^{*m+1}$ for $i \leq N$. We define the usual shift operator $E$ by the relations

$$Ef_i = f_{i+1}$$

$$E^{-1}f_i = f_{i-1},$$

so that

$$(1 + \Delta_x)f_i = Ef_i \tag{3·3·5}$$

$$(1 - \nabla_x)f_i = E^{-1}f_i \tag{3·3·6}$$

from which we may derive the relationship between $\Delta_x$ and $\nabla_x$, namely

$$(1 + \Delta_x) = (1 - \nabla_x)^{-1}$$

If we now expand $(1 - \nabla_x)^{-1}$ we obtain

$$\Delta_x = \nabla_x + \nabla_x^2 + \nabla_x^3 + \dots$$

Using this relationship for $\Delta_x$ in (3·3·3) and (3·3·4), these equations become

$$(3\cdot3\cdot7)\quad \overset{*}{u}_{m+1} = \tfrac{1}{2}(\nabla_x{}^2 + \nabla_x{}^3 + \dots + 2)u_m - ap(2\nabla_x + \nabla_x{}^2 + \nabla_x{}^3 + \dots)f_m$$

$$(3\cdot3\cdot8)\quad u_{m+1} = u_m - {}^p/2\left[\left(1 - \frac{1}{4a}\right)(2\nabla_x + \nabla_x{}^2 + \nabla_x{}^3\dots)f_m + \frac{1}{4a}(2\nabla_x + \nabla_x{}^2 + \nabla_x{}^3 \dots)\overset{*}{f}_{m+1}\right]$$

where (3·3·7) and (3·3·8) are used <u>only</u> for $i = N$. The highest degree of the operator $\nabla_x$ has been purposely omitted as one may include as many terms as one wishes, providing, of course, that there are sufficient points inside the region. By the derivation, formulae (3·3·7) and (3·3·8) have the same truncation error as (3·3·3) and (3·3·4) to order $\gamma$, where we include $\nabla_x{}^\gamma$ as the last term of the $\nabla_x$'s. Hence any error introduced by using this technique differs from that of the interior method only in terms $h^{\gamma+1}$ and higher order terms.

In a similar manner we may introduce the technique for the two dimensional analogue of (3·3·3) and (3·3·4) with the Jacobian matrices having positive eigenvalues once more, namely

$$(3\cdot3\cdot9)\quad \overset{*}{u}_{m+1} = \tfrac{1}{2}(\mu_x + \mu_y)u_m - ap[H_x f_m + H_y g_m]$$

$$(3\cdot3\cdot10)\quad u_{m+1} = u_m - {}^p/2\left[\left(1 - \frac{1}{4a}\right)(H_x f_m + H_y g_m) + \frac{1}{4a}(H_x \overset{*}{f}_{m+1} + H_y \overset{*}{g}_{m+1})\right]$$

There are now three applications to consider when we require

1) $\overset{*}{u}{}^{m+1}_{Nj}$ and $u^{m+1}_{Nj}$ for $j = 1, \dots N-1$
2) $\overset{*}{u}{}^{m+1}_{iN}$ and $u^{m+1}_{iN}$ for $i = 1, \dots N-1$
3) $\overset{*}{u}{}^{m+1}_{NN}$ and $u^{m+1}_{NN}$

Consideration 1) requires the "inversion" (in the sense of the substitution of $\Delta_x$ in terms of $\nabla_x$) of the operator $H_x$ with the operator $H_y$ left alone.

Consideration 2) requires the "inversion" of the operators $H_y$ with the operator $H_x$ left alone.

Consideration 3) requires the "inversion" of both the $H_x$ and $H_y$ operators simultaneously.

We see therefore, using the relationships between the difference operators as shown above, that the schemes for 1), 2) and 3) are

$$1)\quad \overset{*}{u}{}^{m+1}_{Nj}=\tfrac{1}{4}(\nabla_x{}^2+\nabla_x{}^3+\dots+\Delta_y-\nabla_y+4)u^m_{Nj}-ap[2\nabla_x+\nabla_x{}^2+\nabla_x{}^3+\dots)f^m_{Nj}+H_y g^m_{Nj}]$$

$$u^{m+1}_{Nj}=u^m_{Nj}-p/2[(1-\frac{1}{4a})[(2\nabla_x+\nabla_x{}^2+\nabla_x{}^3+\dots)f_{Nj}+H_y g_{Nj}]+\frac{1}{4a}[(2\nabla_x+\nabla_x{}^2+\nabla_x{}^3+\dots)$$

$$\overset{*}{f}{}^{m+1}_{Nj}+H_y\overset{*}{g}{}^{m+1}_{Nj}]]$$

for $j = 1, 2, \dots N-1$

$$2)\quad \overset{*}{u}{}^{m+1}_{1N}=\frac{1}{4}(\Delta_x-\nabla_x+\nabla_y{}^2+\nabla_y{}^3+\dots+4)u^m_{1N}-ap[H_x f^m_{1N}+(2\nabla_y+\nabla_y{}^2+\nabla_y{}^3+\dots)g^m_{1N}]$$

$$u^{m+1}_{1N}=u^m_{1N}-p/2[(1-\frac{1}{4a})[H_x f^m_{1N}+(2\nabla_y+\nabla_y{}^2+\nabla_y{}^3)g^m_{1N}]+\frac{1}{4a}[H_x\overset{*}{f}{}^{m+1}_{1N}+$$

$$(2\nabla_y+\nabla_y{}^2+\nabla_y{}^3+\dots)\overset{*}{g}{}^{m+1}_{1N}]]$$

for $i = 1, 2, \dots N-1$

$$3)\quad \overset{*}{u}{}^{m+1}_{NN}=\frac{1}{4}(\nabla_x{}^2+\nabla_y{}^2+\nabla_x{}^3+\dots+4)u^m_{NN}-ap[(2\nabla_x+\nabla_x{}^2+\nabla_x{}^3+\dots)f^m_{NN}+(2\nabla_y+\nabla_y{}^2+\nabla_y{}^3+\dots)g$$

$$u^{m+1}_{NN}=u_{NN}-p/2[(1-\frac{1}{4a})[(2\nabla_x+\nabla_x{}^2+\nabla_x{}^3+\dots)f^m_{NN}+(2\nabla_y+\nabla_y{}^2+\nabla_y{}^3+\dots)+\frac{1}{4a}[(2\nabla_x+\nabla_x{}^2+\nabla_x{}^3+$$

$$\overset{*}{f}{}^{m+1}_{NN}+(2\nabla_y+\nabla_y{}^2+\nabla_y{}^3+\dots)\overset{*}{g}{}^{m+1}_{NN}]],$$

where theoretical data is used on $x = 0$ and $y = 0$. The corresponding algorithms for the systems of equations in non-conservation form are similarly derived. Their form is obvious and will not be stated.

The above discussion has been for the problem with positive eigenvalues. If the differential problem is one with negative eigenvalues arising from the Jacobian matrix, then data is given along the initial line, as before, but now instead of data given on $x = 0$ (and $y = 0$), it is given on $x = \alpha$ (say) (and $y = \alpha$). The difference schemes are now underdetermined in that data is required by them along the lower boundaries and this data is not forthcoming. Hence a boundary procedure is required for the lower boundaries. A similar analysis to that above will yield such a boundary scheme where, now, the backward difference operators is replaced by a function of the forward difference operators by the relationships,

$$\nabla_x = \Delta_x - \Delta_x^2 + \Delta_x^3 - \Delta_x^4 + \ldots,$$

$$\nabla_y = \Delta_y - \Delta_y^2 + \Delta_y^3 - \Delta_y^4 + \ldots.$$

Similar consideration about the application of this procedure must be taken into account as were considered for the upper boundaries. The above discussion used the general predictor-corrector method as a basis.

This was done merely to show how the method is developed. It can easily be seen that a scheme written in difference form may have the above analysis applied to it and thus a boundary technique developed for the particular method. We do not make this statement general however as later in section (3·5) we will have cause to produce a modified version of this boundary method which can be applied to finite difference procedures in which two simultaneous grid systems operate side by side (for example Friedrich's method [14]), and there is no coupling between adjacent points, see section (3·5).

## 3·4 Application of the boundary technique to the general predictor-corrector method.

We now apply the boundary method introduced in section (3·3) to the general predictor-corrector scheme as it would be used in a problem, that is, in an actual run on the computer. Again considering the one dimensional system of conservation laws

$$\frac{\partial u}{\partial t} + \frac{\partial f}{\partial x} = 0, \; u(x,0) = u_0, \; u(0,t) = u_1$$

with the Jacobian of $f$ with respect to $u$ having positive eigenvalues, the general predictor-corrector method is

$$(3{\cdot}4{\cdot}1) \quad \overset{*}{u}_{m+1} = \mu_x u_m - apH_x f_m$$

$$(3{\cdot}4{\cdot}2) \quad u_{m+1} = u_m - p/2\left[\left(1 - \frac{1}{4a}\right)H_x f_m + \frac{1}{4a} H_x \overset{*}{f}_{m+1}\right].$$

In section (3·3) we left the degree $y$ completely arbitrary. We did not even specify this should be the same at both the predictor level and the corrector level. Since (3·4·1) is a first order approximation to the differential equation, for the boundary procedure to have the same <u>principal part of truncation error</u>, $y$ need only be 2 as then (as may easily be verified) the order $h^2$ truncation error of (3·4·1) and that of

$$(3{\cdot}4{\cdot}3) \quad \overset{*}{u}{}_N^{m+1} = \tfrac{1}{2}(\nabla_x^2 + 2)u_N^m - ap(2\nabla_x + \nabla_x^2)f_N^m$$

are the same. When we consider (3·4·2) however, the combined method is n second order correct and if we require the principal part of the truncation error to be the same between the interior method (3·4·1, 3·4·2) and the boundary method, $y$ must be 3 for (3·4·2). That is the corrector boundary procedure is

$$(3\cdot4\cdot4)\quad u_N^{n+1} = u_N^n - {}^{p}/{}_{2}\left[\left(1-\frac{1}{4a}\right)\left(2\nabla_x + \nabla_x^2 + \nabla_x^3\right)f_N^n + \frac{1}{4a}\left(2\nabla_x + \nabla_x^2 + \nabla_x^3\right)f_N^{*n+1}\right]$$

Thus the actual implementation of the method is as follows:-

1) the predictor formula (3·4·1) is used for $i = 1, 2, \ldots N-1$,

2) the predictor boundary formula (3·4·3) is used for $i = N$,

3) the corrector formula (3·4·2) is used for $i = 1, 2, \ldots N-1$,

4) the corrector boundary formula (3·4·4) is used for $i = N$,

where data from the differential problem is used throughout for $i = 0$.

A similar procedure holds for the two dimensional analogue of (3·4·2) and (3·4·3) where $\gamma$ is 2 for the predictor and 3 for the corrector.

## 3·5 Extrapolation Boundary Techniques involving alternate points only.

As Richtmyer points out, [58] the two step Lax-Wendroff method for the system of conservation laws

$$\frac{\partial u}{\partial t} + \frac{\partial f}{\partial x} = 0$$

yields two simultaneous calculations, one being independent of the other. For example, consider the rectangular mesh of figure (3·5·1) where k and h are the mesh spacings in the time and space directions respectively.

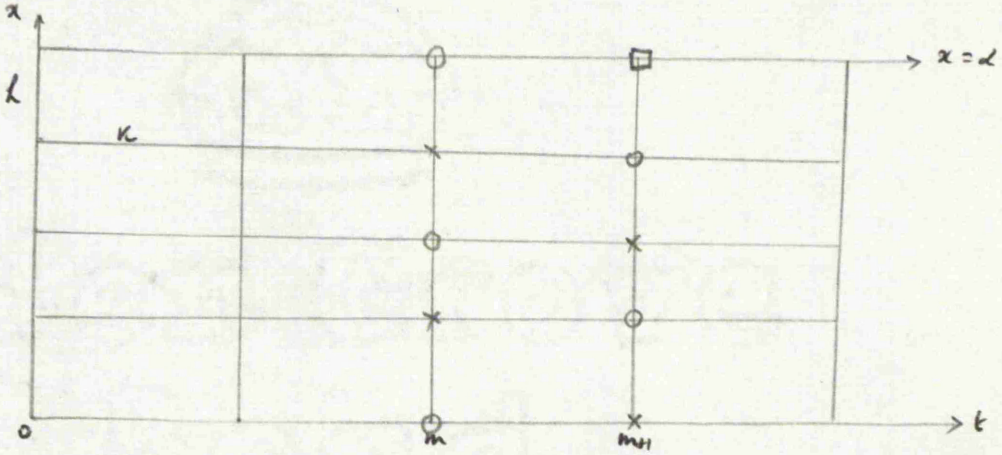


figure (3.5.1)

It may be seen from the two step Lax-Wendroff method (1·6·5, 1·6·6) that the values of u at points at the $(m+1)^{th}$ level marked 0 are determined in terms of values of u and values of f at points at the m'th level marked 0. Similarly values of u at the $(m+1)^{th}$ level at the points marked x are determined in terms of u and f at the m'th level at points marked x, and there is no coupling. Consider the boundary point marked □ at the $(m+1)^{th}$ level. (We are assuming the problem to have a complete set of positive eigenvalues). This point requires a boundary procedure in order that its value be determined. If the straight forward boundary method described in (3·3) is used, then

$$(3\cdot5\cdot1)\quad u^{n+1} = \alpha u_{0_1}^{n} + \beta u_{x_1}^{n} + \delta u_{0_2}^{n} + \epsilon u_{x_2}^{n} + \dots,$$

where $\alpha, \beta, \delta, \epsilon, \dots$ are the coefficients determined by the particular form of the boundary technique derived from the difference operator relations (that is, $\alpha, \beta, , \delta, \epsilon, \dots$ depend on the value of $\gamma$). We have used the notation here, $0_1, 0_2 \dots$ are the nodes numbered from $x = \alpha$, etc Equation (3·5·1) should also include values of the function $f$ but it is sufficient for our purpose just to consider the value of the u's. It can be seen immediately that this boundary method (3·5·1) is imposing a feature which is not present in the difference scheme (1·6·5, 1·6·6). That is, there is a coupling imposed upon the points in the boundary procedure (3·5·1) which is absent in the method (1·6·5, 1·6·6). Hence, we require a boundary procedure which will give us the value of $u$ at $\square$ in terms of all 0's, or all x's, at the previous time level depending on whether $\square$ is an 0 or an x point, respectively. In this section we will develop such a procedure.

Consider the operator $H_x$ acting on $f_i^n$, that is

$$H_x f_i^n = f_{i+1}^n - f_{i-1}^n = -(f_{i-1}^n - f_{i+1}^n).$$

$$(1-\nabla_x)^2 f_{i+1}^n = f_{i-1}^n \quad \text{so that} \quad f_{i+1}^n = (1-\nabla_x)^{-2} f_{i-1}^n, \quad \text{and hence}$$

$$H_x f_i^n = -[1-(1-\nabla_x)^{-2}] f_{i-1}^n.$$

For our boundary technique we require an expression which involves $f$ at $i-1, i-3, i-5, \dots$ at the previous time level as this will then make sure that the boundary point at $\square$ is expressed in terms of 0's <u>or</u> x's. (We have kept the index $i$ in the above representation for convenience. The discussion applies, of course, to a fixed value of

$i = N$). That is we require

$$(3\cdot5\cdot2) \quad [1-(1-\nabla_x)^{-2}]f_{i-1}^{m} = [1+\sum_{j=0}^{N_1} a_j(1-\nabla_x)^{2j}]f_{i-1}^{m},$$

where the $a_j$ are coefficients to be determined and $N_1$ is the accuracy to which we require agreement between the boundary procedure and the interior method. Hence we require, by re-arranging (3·5·2),

$$(3\cdot5\cdot3) \quad -1 = \sum_{j=0}^{N_1} a_j(1-\nabla_x)^{2j+2}$$

This yields the system of equations on equating coefficients of $\nabla_x^{\,j}$,

$$(3\cdot5\cdot4) \quad \sum_{j=0}^{N_1} a_j = -1$$

$$(3\cdot5\cdot5) \quad \sum_{j=0}^{N_1} a_j \binom{2j+2}{m} = 0 \qquad m = 1,2, \ldots N_1$$

where $\binom{2j+2}{m}$ is the usual binomial coefficient, and $m$ corresponds to the coefficients derived from $\nabla_x^{\,m}$ in the expansion of (3·5·3). Thus to obtain the required boundary technique we have to solve (3·5·4) and (3·5·5). To do this we first require a lemma.

**Lemma**

If
$$\sum_{j=0}^{N_1} a_j = -1$$

and
$$\sum_{j=0}^{N_1} \binom{2j+2}{m} a_j = 0 \quad \text{where } \binom{2j+2}{m} = 0 \text{ for } m > 2j+2,\ m = 1,2, \ldots$$

then

$$\sum_{j=0}^{N_1} j^m a_j = (-1)^{m+1} \qquad m = 1, 2, \dots N_1.$$

This lemma may be proved by induction upon m. We now use this lemma to prove the following theorem.

Theorem

The solution of $\sum_{j=0}^{N_1} \binom{2j+2}{m} a_j = 0, \quad m = 1, 2, \dots N_1$

(3.5.8)

and $\sum_{j=0}^{N_1} a_j = -1$ is

$$a_j = \binom{N_1+1}{j+1} (-1)^{j+1}, \; j = 0, 1, \dots N_1$$

where $\binom{N_1+1}{j+1}$ is the usual binomial coeficient.

Proof

Assuming the result of the previous lemma,

$$(3.5.9) \qquad \sum_{j=0}^{N_1} j^m a_j = (-1)^{m+1}$$

The problem of the theorem, therefore, immediately reduces to finding the $a_j$ which satisfy (3.5.9), since these values of $a_j$ will automatically satisfy the problem (3.5.8). If we consider the coefficients defined by the generating function

$$(3.5.10) \quad f(x) = \frac{1}{x}(1-x)^{N_1+1} - \frac{1}{x} = \sum_{i=0}^{N_1} b_i x^i, \qquad x \neq 0$$

we find that on equating coefficients of left and right hand sides,

$$b_i = \binom{N_t+1}{i+1}(-1)^{i+1}$$

To show that the $b_i$ are equivalent to the required $a_j$ we merely have to show that $f(x)$ yields

$$\sum_{i=0}^{N_t} i^m b_i = (-1)^{m+1}.$$

If we put $x = 1$ in (3·5·10),

$$f(x) = -1 = \sum_{i=0}^{N_t} b_i$$

Consider now the derivative with respect to $x$ of (3·5·11). Then

$$f'(x) = \frac{-1}{x^2}(1-x)^{N_t+1} - \frac{(N_t+1)}{x}(1-x)^{N_t} + \frac{1}{x^2} = \sum_{i=0}^{N_t} i b_i x^{i-1}.$$

If we multiply both sides by $x$,

$$-\frac{1}{x}(1-x)^{N_t+1} - (N_t+1)(1-x)^{N_t} + \frac{1}{x} = \sum_{i=0}^{N_t} i b_i x^i.$$

Putting $x = 1$ the left hand side is equal to one and the right hand side yields $\sum_{i=0}^{N_t} i b_i$. By repeating the above procedure it may be seen (by induction if need be) that

$$(3\cdot5\cdot11) \qquad \sum_{i=0}^{N_t} i^m b_i = (-1)^{m+1}.$$

That is the $b_i$ defined by (3·5·10) satisfy the relation (3·5·11) and hence are equivalent to the $a_j$ which we are looking for. Alternatively, the $b_i$ are the solutions to (3·5·10) and are given by

$$b_j = \binom{N_1 + 1}{j + 1}(-1)^{j+1} \qquad j = 0, 1, \ldots N_1$$

and hence the theorem is proved.

We have thus obtained the coefficients for the inversion formula of $H_x$ at the upper boundary. Once more referring to the Lax-Wendroff method (1·6·5, 1·6·6), we see that the difference operator $\mu_x$ also requires 'inversion' at this boundary. It is easy to see that the coefficients in the inversion formula for this operator will be given by finding the expansion equivalent to

$$\tfrac{1}{2}[1 + (1 - \nabla_x)^{-2}]u_{i-1}^{n}$$

to degree $N_1$, that is

$$1 + (1 - \nabla_x)^{-2} = 1 + \sum_{j=0}^{N_1} \tilde{a}_j (1 - \nabla_x)^{2j} \quad ,$$

where the $\tilde{a}_j$ are to be determined. This equation gives rise to the set of equations

$$1 = \sum_{j=0}^{N_1} \tilde{a}_j \;,$$

$$0 = \sum_{j=0}^{N_1} \tilde{a}_j \binom{2j+2}{m} \qquad m = 1, 2, \ldots N_1$$

It may easily be seen that this set of equations has the solution

$$\tilde{a}_j = - a_j \;,$$

where $a_j$ are the coefficients of the 'inversion' formula for $H_x$. If we substitute these coefficients into the respective expansions and substitute the expansion back into the difference scheme, we obtain the boundary replacement

$$u_i^{*m+1} = \tfrac{1}{2}[(1+\tilde{a}_0)u_{i-1}^m + \tilde{a}_1\, u_{i-3}^m + \tilde{a}_2\, u_{i-5}^m + \cdots ] +$$

$$+ ap[(1+a_0)f_{i-1}^m + a_1\, f_{i-3}^m + a_2\, f_{i-5}^m + \cdots ]$$

and

$$u_i^{m+1} = u_i^m + {}^{p}/2[(1-\frac{1}{4a})[(1+a_0)f_{i-1}^m + a_1 f_{i-3}^m + a_2 f_{i-5}^m + \cdots ] +$$

$$+ \frac{1}{4a}[(1-a_0)f_{i-1}^{*m+1} + a_1 f_{i-3}^{*m+1} + a_2 f_{i-5}^{*m+1} + \cdots ]] \quad ,$$

where, once again, it is pointed out that $i = N$ only.

The last terms in the above boundary techniques have necessarily bee omitted. The number of $a_i$'s is determined by the extent of the agreement required between the truncation errors of the boundary technique and the interior region method, or by the extent of the agreeme that is attainable with the number of mesh points available. The extent of the agreement between the boundary replacement and the interior method is necessarily more restricted in this type of boundary modification, than the scheme discussed in (3·3), owing to the fact that we now use only alternate points and hence use twice as many points to attain the same agreement.

## 3·6 Application of boundary techniques to Implicit methods .

So far all discussions related to boundary procedures have been confined to explicit methods. Implicit methods also require boundary data which is not given in the original differential problem. The question therefore arises, can we derive similar schemes for implicit methods? By virtue of their formulation, implicit methods cannot have similar boundary techniques applied to them. However, we note that the predictor formula used in the implicit methods is explicit. We may therefore use the boundary technique similar to the one described in (3·3) for the predictor in order to obtain data at the boundary.

## 3·7 Deferred Approach to the limit.

In the next sections of this chapter, we will consider an important (additional) application of the boundary techniques described earlier in this chapter. This application is to the well known process of deferred approach to the limit. Several authors have considered methods of obtaining improved solutions of differential equations from existing methods by a process of extrapolation. The method of deferred approach to the limit was introduced by Richardson and Gaunt [56] in 1926 for the improved numerical solution of ordinary differential equations. In recent years, authors such as Stetter [63], Gragg [28], Bulirsch and Stoer [4], Mayers (49] and others have extended the ideas of [56] and given the method a more rigorous basis. Most authors, however have confined their attention mainly to the solution of ordinary differential equations. Some, however, have considered certain partial differential equations. In [63] Stetter considers the pure initial value problem for a system of quasilinear hyperbolic equations in two independent variables with two-characteristic directions, and the corresponding extrapolation scheme. The boundary value problems of Neumann and Dirichlet for Laplace's equation and their associated extrapolation schemes are considered by Hofmann [32]. The point worth noting about the examples from partial differential equations is that no boundary technique is required in order to obtain 'extra' boundary data.

As early as 1926, the importance of using 'correct' starting values in linear multistep methods was realised. This concept of 'starting values' has been the topic of investigation, notably by Gragg [27] because the extrapolation procedure depends upon the global truncation error (introduced when discretising the problem) having an asymptotic expansion. (See for example, Gragg [27]). The introduction of 'incorrect' starting values will destroy the simple asymptotic nature of the expansion and hence invalidate the extrapolation. For the 'mid-point' rule for the solution of $y' = f(x,y)$, Gragg devised a starting scheme which maintained the simple asymptotic expansion. In the numerical solution of partial differential equations a different, yet similar, problem exists; namely the incorporation of the extra boundary data required by the difference method. We have seen that the boundary procedure of sections (3·3) and (3·5) have the same expansion as the interior method up to some degree $\gamma$, and hence up to this degree $\gamma$ will have the same asymptotic expansion of the truncation error.

## 3·8 Asymptotic Expansions.

The use of 'deferred approach to the limit' mentioned in the previous section depends on the existence of an asymptotic expansion of the global truncation error. That is, if we have a solution $u_i^n$ of the difference equation and a smooth solution $U$ of the differential equation at the same point $(x, t)$ then

$$U - u_i^n = \sum_{i=1}^{\infty} C_i h^i = D(h), \quad (\text{since } k = ph,\ p\ .\ \text{constant})$$

where $C_i$ are functions independent of $h$. The method of extrapolation is then to take successive approximations to $D(h)$ by a series of $h_i$, $i = 0, 1, 2, \ldots$ where

$$h_0 > h_1 > h_2 > \ldots$$

and to eliminate the successive powers of $h$ in $D(h)$ by taking suitable combinations of the approximations.

The proof of the existence of the asymptotic expansion of a particular finite difference method, and the derivation of its actual form, is obviously a very complex task. Gragg's thesis [27] was devoted to the proof of such expansions for problems in ordinary differential equations, the proof here will obviously be even more difficult. We do not try to indicate how this may be done, although presumably, with suitable conditions upon the functions dealt with, an analysis along the lines given by Gragg would yield some results.

In the application we consider here, an expansion of the form

$$(3{\cdot}8{\cdot}1)\quad D(h) = C_1(x,t)h + C_2(x,t)h^2 + C_3(x,t)h^3 + \ldots$$

will be assumed where $C_1$, $C_2$, $C_3$, ... are functions of $x$ and $t$.

We will then determine, along similar lines to those of Mayers [49], whether or not there is an order h term in $D(h)$, that is, whether or not $C_1 = 0$. Consider the general predictor-corrector method for the system of conservation laws

(3·8·2) $\frac{\partial u}{\partial t} + \frac{\partial f}{\partial x} = 0,$

namely

(3·8·3) $\overset{*}{u}_{m+1} = \mu_x u_i^m - apH_x f_i^m$

(3·8·4) $u_{m+1} = u_m - {}^p/_2[(1 - \frac{1}{4a})H_x f_i^m + \frac{1}{4a}H_x \overset{*}{f}{}_i^{m+1}].$

If we assume an expansion of the form (3·8·1), we require to determine the $C_i$ from (3·8·2, 3·8·3, 3·8·4). Eliminate the $*_{m+1}$ level between (3·8·3) and (3·8·4) and

(3·8·5) $u_{m+1} = u_m - {}^p/_2[(1-\frac{1}{4a})H_x f_i^m + \frac{1}{4a} H_x f(\mu_x u_i^m - apH_x f_i^m)]$

is obtained. Expanding (3·8·5) in terms of $u_m$ and its derivatives using Taylor's series and (3·8·2), leads to

$$[k\frac{\partial}{\partial t} + \frac{k^2}{2}\ \frac{\partial^2}{\partial t^2} + \dots]u_i^m = - ph\frac{\partial f}{\partial x} + \frac{p^2h^2}{2}\frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\frac{\partial f}{\partial u}\right) + 0(h^3)$$

If we substitute for $u_i^m$ from (3·8·1), then we obtain

$$[k\frac{\partial}{\partial t} + \frac{k^2}{2}\frac{\partial^2}{\partial t^2} + \dots][U - C_1h - C_2h^2 + \dots] = -ph\frac{\partial f}{\partial x}(U - C_1h - C_2h^2 + \dots) + + \frac{p^2h^2}{2}\left(\frac{\partial f}{\partial x}(U) \cdot \frac{\partial f}{\partial u}(U)\right) + 0(h^3)\ .$$

Since this is an identity, we may equate coefficients of powers of h, giving

(3·8·6) $k\frac{\partial U}{\partial t} = - ph\frac{\partial f}{\partial x}(U),$

(3·8·7) $\frac{k^2}{2}\frac{\partial^2 U}{\partial t^2} - kh\frac{\partial C_1}{\partial t} = kh\frac{\partial}{\partial x}\left(C_1\frac{\partial f}{\partial u}\right) + \frac{k^2}{2}\frac{\partial}{\partial x}\left[\frac{\partial f(U)}{\partial x} - \frac{\partial f(U)}{\partial u}\right]$ .

Result (3·8·6) merely states that $U$ should be the solution of the differential equation. From (3·8·2) and since $\frac{\partial}{\partial U} = \frac{\partial}{\partial u}$, it follows that

$$\frac{\partial^2 U}{\partial t^2} = -\frac{\partial}{\partial x}\left(\frac{\partial f}{\partial U}\frac{\partial U}{\partial t}\right) = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial U}\frac{\partial f}{\partial x}\right).$$

Hence (3·8·7) reduces to

(3·8·8) $\frac{\partial C_1}{\partial t} + \frac{\partial}{\partial x}\left(C_1\frac{\partial f}{\partial U}\right) = 0$ with $C_1(x,0) = C_1(0,t) = 0$.

The solution of (3·8·8), which is a homogeneous problem is $C_1 = 0$ provided $\frac{\partial f}{\partial u}\left(=\frac{\partial f}{\partial U}\right)$ is a 'well-behaved' function. A similar analysis may be given to show that $C_2 \neq 0$, $C_3 \neq 0$, etc. Hence we see that $C_1 = 0$ implies that there will be no term of order $h$ in the asymptotic expansion. We therefore use the general predictor-corrector method and extrapolate with an expansion of the form

(3·8·9) $C_2h^2 + C_3h^3 + \ldots$

and the $O(h^2)$ and $O(h^3)$ ... terms may be eliminated by simple linear elimination.

## 3·9 Numerical Results.

In this section we shall present the results of a number of numerical experiments carried out using the material presented in this chapter. The first set of experiments was devised to consider the effect of iteration of the corrector described in section (3·1) for various values of the parameter 'a'. The problem considered was the one dimensional system

$$(3\cdot9\cdot1) \quad \frac{\partial u}{\partial t} + \frac{\partial}{\partial x}(\tfrac{1}{2}u^2) = 0 \quad \begin{array}{ll} u(x,0) = x & 0 \leq x \leq 1 \\ u(0,t) = 0 & t \geq 0 \end{array}$$

which has the theoretical solution

$$u(x,t) = \frac{x}{1+t} .$$

The results using the general predictor-corrector scheme of Chapter 2 as a means of solution are shown in table (3·9·1).

The next set of experiments considered the same problem as above but now used the boundary replacement discussed in section (3·3) where $y = 2$ for predictor level and $y = 3$ for the corrector level. The results are shown for (3·9·1) in table (3·9·2).

We next considered the application of the boundary technique of (3·3) to the deferred approach to the limit discussed in section (3·8) for (3·9·1). The value of $y$ was taken to be $6$ for this problem in order that the truncation errors of the boundary scheme and the interior method were the same up to order $h^6$. The results are given in table (3·9·3) where two and three extrapolations are shown. The result of taking theoretical data at the upper boundary is shown in table (3·9·4

## Tables of Results

Table (3·9·1) The one dimensional problem (3·9·1)

| p | Number of time steps | $a=\frac{1}{4}$ | | $a=\frac{1}{2}$ | | $a=1$ | |
|---|---|---|---|---|---|---|---|
| Number of Iterations | | 0 | 2 | 0 | 2 | 0 | 2 |
| 0·6 | 100 | $\cdot 15\times 10^{-4}$ | $\cdot 19\times 10^{-1}$ | $\cdot 17\times 10^{-3}$ | $\cdot 35\times 10^{-3}$ | $\cdot 37\times 10^{-2}$ | $\cdot 10\times 10^{-1}$ |
| | 300 | $\cdot 11\times 10^{-2}$ | $\cdot 23\times 10^{-1}$ | $\cdot 85\times 10^{-3}$ | $\cdot 20\times 10^{-4}$ | $\cdot 11\times 10^{-3}$ | $\cdot 79\times 10^{-2}$ |
| 1·0 | 100 | $\cdot 11\times 10^{-2}$ | $\cdot 29\times 10^{-1}$ | $\cdot 11\times 10^{-2}$ | $\cdot 25\times 10^{-3}$ | $\cdot 57\times 10^{-3}$ | $\cdot 92\times 10^{-2}$ |
| | 300 | $\cdot 36\times 10^{-3}$ | * | $\cdot 22\times 10^{-2}$ | $\cdot 61\times 10^{-3}$ | $\cdot 25\times 10^{-2}$ | * |

We have included the results of the general predictor-corrector method as given in Chapter 2 (0 - iterations) in order to give a comparison between the results without and with iteration. The entries marked * indicate that non-linear instability had developed. The mesh size h was 0·1.

<u>Table 3·9·2</u> Problem (3·9·1) solved by the general predictor-corrector and boundary replacement of (3·3) ($y=2$, $y=3$ for predictor and corrector respectively).

| 'a' | p | Errors across the region: every second point | | | | |
|---|---|---|---|---|---|---|
| | 0·3 | $\cdot 33 \times 10^{-5}$ | $\cdot 99 \times 10^{-5}$ | $\cdot 16 \times 10^{-4}$ | $\cdot 23 \times 10^{-4}$ | $\cdot 29 \times 10^{-4}$ |
| 0·25 | 0·6 | $\cdot 50 \times 10^{-5}$ | $\cdot 15 \times 10^{-4}$ | $\cdot 25 \times 10^{-4}$ | $\cdot 35 \times 10^{-4}$ | $\cdot 45 \times 10^{-4}$ |
| | 1·0 | $\cdot 62 \times 10^{-5}$ | $\cdot 18 \times 10^{-4}$ | $\cdot 31 \times 10^{-4}$ | $\cdot 43 \times 10^{-4}$ | $\cdot 55 \times 10^{-4}$ |
| | 0·3 | $\cdot 21 \times 10^{-5}$ | $\cdot 65 \times 10^{-5}$ | $\cdot 11 \times 10^{-4}$ | $\cdot 15 \times 10^{-4}$ | $\cdot 19 \times 10^{-4}$ |
| 0·5 | 0·6 | $\cdot 33 \times 10^{-5}$ | $\cdot 99 \times 10^{-5}$ | $\cdot 16 \times 10^{-4}$ | $\cdot 23 \times 10^{-4}$ | $\cdot 29 \times 10^{-4}$ |
| | 1·0 | $\cdot 39 \times 10^{-5}$ | $\cdot 11 \times 10^{-4}$ | $\cdot 19 \times 10^{-4}$ | $\cdot 27 \times 10^{-4}$ | $\cdot 35 \times 10^{-4}$ |
| | 0·3 | $\cdot 10 \times 10^{-6}$ | $\cdot 17 \times 10^{-6}$ | $\cdot 16 \times 10^{-6}$ | $\cdot 3 \times 10^{-6}$ | $\cdot 44 \times 10^{-6}$ |
| 1·0 | 0·6 | $\cdot 21 \times 10^{-6}$ | $\cdot 72 \times 10^{-6}$ | $\cdot 11 \times 10^{-5}$ | $\cdot 14 \times 10^{-5}$ | $\cdot 18 \times 10^{-5}$ |
| | 1·0 | $\cdot 47 \times 10^{-6}$ | $\cdot 14 \times 10^{-5}$ | $\cdot 24 \times 10^{-5}$ | $\cdot 34 \times 10^{-5}$ | $\cdot 43 \times 10^{-5}$ |

The mesh size $h$ was again taken to be 0·1.

**Table 3·9·3** Problem (3·9·1) with general predictor-corrector solution and boundary replacement of (3·3), $y = 6$. The asymptotic expansion was assumed to be of the form (3·8·9); the extrapolation was performed by simple elimination of the terms.

ERRORS at $100 \times p \times h_0$, at $x = \frac{1}{2}$

| a | p | $h_0 = 0{\cdot}1$ | $h_1 = 0{\cdot}05$ | $h_2 = 0{\cdot}025$ | $h_3 = 0{\cdot}0125$ | 2 Exts. | 3 Exts. |
|---|---|---|---|---|---|---|---|
|  | 0·3 | ·13 x 10-4 | ·10 x 10-5 | ·79 x 10-6 | ·19 x 10-6 | ·19 x 10-9 | ·18 x 10-12 |
| 0·25 | 0·6 | ·20 x 10-4 | ·45 x 10-5 | ·11 x 10-5 | ·29 x 10-6 | ·10 x 10-8 | ·18 x 10-11 |
|  | 1·0 | ·34 x 10-4 | ·59 x 10-5 | ·14 x 10-5 | ·80 x 10-6 | ·52 x 10-8 | ·78 x 10-11 |
|  | 0·3 | ·85 x 10-5 | ·21 x 10-5 | ·52 x 10-6 | ·13 x 10-6 | ·41 x 10-11 | ·16 x 10-12 |
| 0·5 | 0·6 | ·13 x 10-4 | ·34 x 10-5 | ·79 x 10-4 | ·19 x 10-6 | ·81 x 10-11 | ·17 x 10-11 |
|  | 1·0 | ·15 x 10-4 | ·38 x 10-5 | ·95 x 10-6 | ·23 x 10-6 | ·18 x 10-9 | ·10 x 10-10 |

Exts. = extrapolations

**Table 3·9·4:** As for table (3·9·3) but with theoretical boundary data

ERRORS at 100 x p x $h_0$ at $x = \frac{1}{2}$

| a | p | $h_0$ = 0·1 | $h_1$ = 0·05 | $h_2$ = 0·025 | $h_3$=0·125 | 2 Exts. | 3 Exts. |
|---|---|---|---|---|---|---|---|
| | 0·3 | ·63 x 10-4 | ·17 x 10-4 | ·17 x 10-5 | ·46 x 10-7 | ·15 x 10-4 | ·24 x 10-5 |
| 0·25 | 0·6 | ·15 x 10-4 | ·39 x 10-4 | ·29 x 10-5 | ·10 x 10-6 | ·25 x 10-4 | ·52 x 10-5 |
| | 1·0 | ·11 x 10-2 | ·43 x 10-4 | ·39 x 10-5 | ·12 x 10-6 | ·26 x 10-4 | ·24 x 10-5 |
| | 0·3 | ·13 x 10-3 | ·28 x 10-4 | ·17 x 10-5 | ·63 x 10-7 | ·72 x 10-5 | ·96 x 10-6 |
| 0·5 | 0·6 | ·17 x 10-3 | ·21 x 10-4 | ·99 x 10-5 | ·14 x 10-6 | ·51 x 10-5 | ·48 x 10-5 |
| | 1·0 | ·11 x 10-2 | ·34 x 10-3 | ·30 x 10-5 | ·21 x 10-6 | ·14 x 10-3 | ·24 x 10-4 |

Exts. = extrapolations

## 3·10 Comments on Chapter 3.

From the numerical results we can see that in order to use the iterative corrector, the correct value of $a = \frac{1}{2}$ must be used. The results with $a = \frac{1}{4}$ and $1$ yield non-linear instability in some cases and in the cases where it does not arise, the accuracy is not as good as the results with $a = \frac{1}{2}$. In practice it was found that 2 iterations were sufficient for convergence. The improvement in the solutions obtained as a result of using the boundary replacement as can be seen from table (3·11·2) is quite considerable.

There are three reasons why we believe the boundary replacement method is more acceptable than iteration of the corrector.

First, the problem is no longer over-determined since the extra data required by the difference method is generated by the boundary procedure. The iterative method still applies to overdetermined differential problems.

Second, the boundary method requires considerably less work and time than does the iterative method. This is, of course, particularly important in problems in two and more space dimensions.

Third, we feel justified in saying that the results obtained as a result of using the boundary technique are more accurate than those obtained from the iterative method.

Finally, the last tables show the considerable accuracy that is achievable with a quite modest number of extrapolations. They also show how very carefully data for an under determined difference problem must be added in order that the asymptotic expansions be maintained. In this chapter we have endeavoured to derive techniques for finding solutions to underdetermined partial difference systems which appear to work well in certain circumstances. We do not attempt to answer, here, the difficulties on theoretical grounds, but it is hoped that the insight gained from the experiments will lead to a firmly based extrapolation procedure (on the lines of the work of Gragg [27]) for first order hyperbolic systems.

The procedures outlined can be extended in a natural way to systems in many space dimensions and can also be applied to develop extrapolation schemes for other underdetermined problems.

# CHAPTER 4

# EXTENSIONS TO GENERAL QUASILINEAR DIFFERENTIAL SYSTEMS

GOURLAY AND MORRIS [25]

## Introduction.

In Chapter 2 we considered numerical methods of solving the systems of partial differential equations which were either systems of conservation laws, or were quasilinear and homogeneous, that is, the coefficient matrix of the first derivative in space was a function only of the components of the unknown variable u, and there were no vector functions of u, x, y and t appearing on the right hand sides of the equations. In this chapter, we extend the methods derived in Chapter 2 to solve the general quasilinear system of partial differential equations

$$\frac{\partial u}{\partial t} + \frac{\partial f(u, x, t)}{\partial x} = \gamma(u, x, t),$$

and

$$\frac{\partial u}{\partial t} + A(u, x, t)\frac{\partial u}{\partial x} = \gamma(u, x, t),$$

together with their corresponding two dimensional analogues.

Systems of conservation laws, as was seen in Chapter 1, arise as a result of neglecting dissipative phenomena such as viscosity and friction. We would like to be able to generate finite difference methods of high accuracy which remain stable and which are efficient to use for the solution of systems of partial differential equations of the form written above, and for their two dimensional analogues. These equations occur naturally as a result of including the dissipative terms mentioned above.

A class of problems which falls into this category involves the transport equations in neutron diffusion and radiative transfer. These equations (which for brevity we do not state) are dealt with numerically by Richtmyer [57], Keller and Wendroff [37], Carlson [7], and others. Another class of problems which may be written in the above form contains the Navier-Stoke's Equations. The numerical solution of these equations has been the subject of papers by such authors as Thommen [66], [67], Gary [18] and Crocco [11].

In order to formulate finite difference methods for solving these equations, we require a notation which is consistent with that used in the previous chapters but which allows the inclusion of additional terms. We again superimpose a rectangular grid on the region of computation of mesh size h in the space co-ordinates and k in the time co-ordinate and let the ratio k/h = p, a constant. The values of the space and time variables are represented by

$$x = ih, \quad y = jh, \quad t = mk \quad \text{for } i, j = 0, 1, 2, \dots N \text{ and}$$
$$m = 0, 1, 2, \dots$$

and then

$$u(x, y, t) \equiv u(ih, jh, mk) \qquad (= u_m)$$

The difference operators $H_x$, $H_y$, $\mu_x$, $\mu_y$ are defined as before and the additional notation involves the functions f, g and $\gamma$ where

$$f_m = f_i^m = f(u_i^m, ih, mk)$$
$$f^*_{m+1} = f_i^{*m+1} = f(\overset{*}{u}{}_i^{m+1}, ih, mk)$$
$$\bar{f}_{m+1} = \bar{f}_i^{m+1} = f(u_i^m, ih, (m+1)k),$$

and for the two dimensional problems

$$f_m \equiv f^m_{ij} \equiv f(u^m_{ij}, ih, jh, mk)$$

$$f^*_{m+1} \equiv f^{*m+1}_{ij} \equiv f(u^{*m+1}_{ij}, ih, jh, mk)$$

$$\bar{f}_{m+1} \equiv \bar{f}^{m+1}_{ij} \equiv f(u^m_{ij}, ih, jh, (m+1)k),$$

and similar notations for $g$ and $\gamma$ in one and two dimensions.

Other notations which are used only in particular cases are defined at the appropriate places.

## 4.1 Explicit One Dimensional case.

Consider the first order hyperbolic system

$$(4.1.1) \quad \frac{\partial u}{\partial t} + \frac{\partial f}{\partial x}(u, x, t) = \gamma_1(u, x, t), \quad \begin{cases} u(x, o) = u_o(x) & x \geq o \\ u(o, t) = u_1(t) & t \geq o \end{cases}$$

where u is an n-component unknown vector function of x and t, and f and $\gamma$ are known vector functions of the components of u, and of x and t. The solution of this problem will be required in the region

$$o \leq x \leq \alpha, \; t \geq o \quad \text{where } Nh = \alpha.$$

Initial data $u(x, o) = u_o(x)$ given for $x \geq o$ will be assumed although only that part of $u_o(x)$ for $o \leq x \leq \alpha$ will be required. If we assume that the Jacobian matrix of $f(u, x, t)$ with respect to the components of u has a complete set of positive eigenvalues, then (4.1.1) together with the initial data $u_o(x)$ constitutes a properly posed problem only if data is also given on $x = o$, $t \geq o$; that is $u(o, t) = u_1(t)$, $t \geq o$. The problem with a complete set of positive eigenvalues has been considered only to simplify the analysis. The analysis, however, may be carried through in more general cases.

If the differentiation in (4.1.1) is carried out, then

$$(4.1.2) \qquad \frac{\partial u}{\partial t} + A(u, x, t) \frac{\partial u}{\partial x} = \gamma_2(u, x, t),$$

is obtained where $A(u, x, t)$ has a complete set of positive eigenvalues. Whilst (4.1.2) may always be obtained from (4.1.1) (provided, of course, the derivatives exist), the process may not always be reversible. That is some systems of the form (4.1.2) may either prove impossible or extremely difficult to write in the form (4.1.1). For this reason finite difference schemes for (4.1.2) are considered separately from those for (4.1.1). The explicit methods derived will again be general predictor-corrector methods which bear a strong resemblance to the general predictor-corrector methods considered in Chapter 2. The methods derived in this chapter will be such that when the systems (4.1.1) and (4.1.2) reduce to systems

$$(4.1.3) \qquad \frac{\partial u}{\partial t} + \frac{\partial f}{\partial x} = 0$$

$$(4.1.4) \qquad \frac{\partial u}{\partial t} + A(u)\frac{\partial u}{\partial x} = 0$$

considered in Chapter 2, they reduce to the general predictor-corrector methods of Chapter 2.

Thus consider the explicit predictor-corrector method for the solution of (4.1.1), (with $\mathfrak{z} \equiv \mathfrak{z}_1$), namely

$$(4.1.5) \qquad u^*_{m+1} = \mu_x u_m - ap(H_x f^m_i - 2h\mathfrak{z}^m_i)$$

$$(4.1.6) \qquad u_{m+1} = u_m - p[bH_x f^m_i + cH_x f^{*m+1}_i + eH_x \bar{f}^{m+1}_i$$

$$+ 2h(d\mathfrak{z}^{*m+1}_i + s\bar{\mathfrak{z}}^{m+1}_i + q\mathfrak{z}^m_i)],$$

where the notation proposed in the introduction of this chapter has been used, and where a, b, c, d, e, s, q are constants to be determined. Once again (c.f. Chapter 2) equation (4.1.5) is termed the predictor. It is a first order approximation to (4.1.1) located at the time level $(m + 2a)k$. We require that the overall scheme (4.1.5, 4.1.6) obtained as a result of eliminating $u^*_{m+1}$ from (4.1.6) be a second order correct approximation to (4.1.1). Carrying out the elimination of $u^*_{m+1}$ between (4.1.5) and (4.1.6), it may be seen that the resulting expression is of the form

$$(4.1.7)\quad u_{m+1} = u_m - p[bH_x f_i^m + cH_x f(\delta_x u_m - ap(H_x f_i^m - 2h\gamma_i^m)) + eH_x f_i^{m+1} +$$

$$+ 2h[d\gamma(\delta_x u_m - ap(H_x f_i^m - 2h\gamma_i^m)) + s\gamma_i^{m+1} + q\gamma_i^m]]$$

Expanding (4.1.7) first using Taylor series and retaining terms up to and including those of order $h^2$, then substituting $H_x \doteq 2h\frac{\partial}{\partial x} + O(h^3)$, the result

$$(4.1.8)\quad u_{m+1} = u_m - 2ph[(b + c + e)\frac{\partial f}{\partial x} + (d + s + q)\gamma]_i^m +$$

$$+ p^2h^2[4ac\frac{\partial}{\partial x}(A(\frac{\partial f}{\partial x} - \gamma)) + 4ad(\frac{\partial f}{\partial x} - \gamma)\frac{\partial \gamma}{\partial u} - 2e\frac{\partial}{\partial x}(\frac{\partial f}{\partial t}) -$$

$$- 2s\frac{\partial \gamma}{\partial t}]_i^m + O(h^3),$$

is obtained. An expansion of $u_i^{m+1}$ in terms of $u_i^m$ and its derivatives using Taylor's series yields

$$(4.1.9)\quad u_{m+1} = u_m - ph\left(\gamma - \frac{\partial f}{\partial x}\right)_i^m + \frac{p^2h^2}{2}\left[\frac{\partial \gamma}{\partial t} + \frac{\partial \gamma}{\partial u}\left(\gamma - \frac{\partial f}{\partial x}\right) - \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial t} + A\left(\gamma - \frac{\partial f}{\partial x}\right)\right)\right]_i^m + O(h^3),$$

where use has once again been made of (4.1.1).

Comparing the coefficients of the respective terms in (4.1.8) and (4.1.9) we see that

$$2(b + c + e) = 1$$
$$2(d + s + q) = -1$$
$$4ac = \tfrac{1}{2}$$
$$4ad = -\tfrac{1}{2}$$
$$2e = \tfrac{1}{2}$$
$$2s = -\tfrac{1}{2},$$

must be satisfied in order that (4.1.5), 4.1.6) be second order correct. On solving these equations it is found that

$$c = -d = \frac{1}{8a}, \qquad b = -q = \frac{1}{4}\left(1 - \frac{1}{2a}\right), \qquad e = -s = \frac{1}{4}$$

where 'a' is a parameter, so that the required difference method is

$$u^*_{m+1} = \mu_x u_m - ap(H_x f_i^m - 2h\gamma_i^m)$$

(4.1.10)

$$u_{m+1} = u_m - \frac{p}{4}\left[\left(1 - \frac{1}{2a}\right)(H_x f_i^m - 2h\gamma_i^m) + \frac{1}{2a}(H_x f_i^{*m+1} - 2h\gamma_i^m) + \right.$$

$$\left. + (H_x \bar{f}_i^{m+1} - 2h\bar{\gamma}_i^{m+1})\right]$$

For the system of equations (4.1.2) the similar scheme (with $\zeta \equiv \zeta_2$)

(4.1.11)

$$u^{*}_{m+1} = \mu_x u_m - ap(A_m H_x u_i^m - 2h_{\zeta i}^{m})$$

$$u_{m+1} = u_m - \frac{p}{4}\Big[(1 - \frac{1}{2a})(A_m H_x u_i^m - 2h_{\zeta i}^{m}) + \frac{1}{2a}(A^{*}_{m+1} H_x u_i^{*m+1} - 2h_{\zeta i}^{*m+1})$$

$$+ (\bar{A}_{m+1} H_x u_i^m - 2h_{\zeta i}^{\overline{m+1}})\Big]$$

where

$$A_m \equiv A_i^m \equiv A(u_i^m, x, t_m),\ A^{*}_{m+1} = A^{*m+1}_i = A(u^{*m+1}_i, x, t_m),$$

$$\bar{A}_{m+1} \equiv \bar{A}_i^{m+1} \equiv A(u_i^m, x, t_{m+1})$$

is obtained. It may be noted that the generalizations of the predictor corrector methods of Chapter 2, that is (4.1.10) and (4.1.11) still have a parameter 'a' defining the location of the predictor level in time. An interesting point to note is that when $a = \frac{1}{4}$, (4.1.10) and (4.1.11) reduce to formulae which are generalizations of the two step Lax-Wendroff method for the system of general quasilinear partial differential equations considered here.

## 4.2 Generalizations of the Boundary Techniques.

If we continue to consider the system of general quasilinear partial differential equations (4.1.1) and (4.1.2) where the Jacobian matrix $A(u, x, t)$ has a complete set of positive eigenvalues, in order to solve the difference schemes derived in the previous chapter a technique is required for introducing data on the line $x = \alpha$, $t>0$. As was seen in Chapter 3 there were two approaches to this problem. First we could insert numerical data on $x = \alpha$, for example theoretical values along $x = \alpha$, and then use an iterative process, and second using another difference replacement at $x = \alpha$ which is consistent (that is, has the same truncation error up to some pre-assigned degree) with the difference method used in the interior of the region. The boundary replacement for (4.1.1) and (4.1.2) is now given.

Consider first the generalization of the iterative process (3.1.4). Consider the general predictor-corrector method derived in section (4.1) for the solution of (4.1.1), namely

$$u^{*}_{m+1} = \mu_x u_m - ap(H_x f_i^{m} - 2h\mathfrak{z}_i^{m}) \tag{4.2.1}$$

$$u_{m+1} = u_m - \frac{p}{4}[(1 - 1/2a)(H_x f_i^{m} - 2h\mathfrak{z}_i^{m}) + \frac{1}{2a}(H_x f_i^{*m+1} - 2h\mathfrak{z}_i^{*m+1}) + (H_x \bar{f}_i^{m+1} - 2h\bar{\mathfrak{z}}_i^{m+1})] \tag{4.2.2}$$

Equations (4.2.1) and (4.2.2) may be regarded as generating the predictor corrector scheme

$$(4.2.3) \qquad u^{*}_{m+1} = \mu_x u_m - ap(H_x f_i^m - 2h\gamma_i^m)$$

$$(4.2.4) \qquad u^{(j+1)}_{m+1} = u_m - \frac{p}{4}[(1 - \frac{1}{2a})(H_x f_i^m - 2h\gamma_i^m) + \frac{1}{2a}(H_x f_i^{(j)m+1}$$

$$- 2h\gamma_i^{(j)m+1}) + (H_x \bar{f}_i^{m+1} - 2h\bar{\gamma}_i^{m+1})] \quad j = 1, 2, \ldots ,$$

where

$$f_i^{(1)m+1} \doteq f_i^{*m+1}, \qquad \gamma_i^{(1)m+1} \doteq \gamma_i^{*m+1}.$$

We are really therefore finding the solution of the replacement

$$(4.2.5) \qquad u_{m+1} = u_m - \frac{p}{4}[(1 - \frac{1}{2a})(H_x f_i^m - 2h\gamma_i^m) + \frac{1}{2a}(H_x f_i^{m+1} - 2h\gamma_i^{m+1})$$

$$+ (H_x \bar{f}_i^{m+1} - 2h\bar{\gamma}_i^{m+1})$$

by an iterative process. Expanding the right hand side of (4.2.5) in terms of $u_m$ and its derivatives using Taylor's series and retaining terms up to and including those of order $h^2$, we obtain

$$(4.2.6) \quad u_{m+1} = [u + k(\gamma - \frac{\partial f}{\partial x}) + \frac{k^2}{2}\{\frac{\partial \gamma}{\partial t} - \frac{\partial}{\partial x}(\frac{\partial f}{\partial t}) + \frac{1}{2a}(\frac{\partial}{\partial x}(\frac{\partial f}{\partial x} - \gamma)\frac{\partial f}{\partial u}) -$$

$$- (\frac{\partial f}{\partial x} - \gamma)\frac{\partial \gamma}{\partial u}\}]_m + O(h^3),$$

whilst an expansion of $u_{m+1}$ in terms of $u_m$ and its derivatives by Taylor's series yields

$$(4.2.7)\quad u_{m+1} = \Big[u + k\Big(\gamma - \frac{\partial f}{\partial x}\Big) + \frac{k^2}{2}\Big\{\frac{\partial \gamma}{\partial t} - \frac{\partial}{\partial x}\Big(\frac{\partial f}{\partial t}\Big) - \frac{\partial}{\partial x}\Big(\Big(\gamma - \frac{\partial f}{\partial x}\Big)\frac{\partial f}{\partial u}\Big) - \Big(\frac{\partial f}{\partial x} - \gamma\Big)\frac{\partial \gamma}{\partial u}\Big\}\Big]_m + 0(h^3),$$

where in both (4.2.6) and (4.2.7), result (4.1.1) has been used. Comparing (4.2.6) and (4.2.7) we see that whilst (4.2.6) is a consistent replacement of (4.2.7) for general 'a', only the choice $a = \frac{1}{2}$ gives a corrector which <u>may</u> converge to the order of the overall scheme (4.2.4, 4.2.5). That is, the method with $a = \frac{1}{2}$, provided certain convergence conditions (see later) are satisfied, will converge to the solution of the differential problem in such a way that the principal part of the truncation error is of order $h^3$, whereas when $a = \frac{1}{4}$, the method will converge to the solution of the differential equation with the principal part of the truncation error of order $h^2$. Thus the iterative scheme takes the form (on substituting $a = \frac{1}{2}$)

$$(4.2.8)\quad u^*_{m+1} = \mu_x u_m - \frac{p}{2}[H_x f_i^m - 2h\gamma_i^m]$$

$$(4.2.9)\quad u^{(j+1)}_{m+1} = u_m - \frac{p}{4}[(H_x f_i^{(j)m+1} - 2h\gamma_i^{(j)m+1}) + (H_x f_i^{-m+1} - 2h\zeta_i^{-m+1})].$$

A similar iterative scheme may be derived for the system of equations (4.1.2), namely

$$u^*_{m+1} = \mu_x u_m - \frac{p}{2}[A_m H_x u_m - 2h\zeta_i^m]$$

$$u_{m+1}^{(j+1)} = u_m - \frac{p}{4}[A_{m+1}^{(j)} H_x u_{m+1}^{(j)} - 2h_{m+1}^{(j)} + \bar{A}_{m+1} H_x u_m - 2\bar{h}_{m+1}] ,$$

where $A_{m+1}^{(j)} \doteq A(u_i^{(j)m+1})$, and $a = \frac{1}{2}$

The conditions for the convergence of these schemes will be discussed in (4.8).

As was pointed out in Chapter 3, the iteration of the corrector smoothes out the error introduced by substituting theoretical data on $x = X$; it <u>does not</u> eliminate altogether the errors thus introduced. We therefore prefer a replacement which will avoid introducing these "inconsistent" errors. As was seen in Chapter 3 for the simpler equations, we were able to derive a boundary replacement. This idea will now be extended to the system of equations considered in this chapter.

Consider (4.1.10) for the solution of the problem with positive eigenvalues. Following the analysis of Chapter 3, we use the relationships, derived there, between the difference operators so that $\Delta_x = \nabla_x + \nabla_x^2 + \nabla_x^3 + \ldots$ . Eliminating $\Delta_x$ from the operator $H_x$ and $\mu_x$, then

$$H_x \rightarrow (2\nabla_x + \nabla_x^2 + \nabla_x^3 + \ldots) \text{ and } \mu_x \rightarrow \tfrac{1}{2}(\nabla_x^2 + \nabla_x^3 + \ldots + 2).$$

The boundary replacement therefore becomes:

$$(4.2.10) \quad u_{m+1}^{*} = \tfrac{1}{2}(\nabla_x^2 + \nabla_x^3 + \ldots + 2)u_m - ap[(2\nabla_x + \nabla_x^2 + \nabla_x^3 + \ldots)f_i^m - 2h_i^m]$$

$$(4.2.11)\quad u_{m+1} = u_m - \frac{p}{4}\Big[\Big(1 - \frac{1}{2a}\Big)\Big\{(2\nabla_x + \nabla_x^{\,2} + \nabla_x^{\,3} + \dots)f_i^m - 2h\,g_i^m\Big\} +$$

$$+ \frac{1}{2a}\Big\{(2\nabla_x + \nabla_x^{\,2} + \nabla_x^{\,3} + \dots)f_i^{*m+1} - 2h\,g_i^{*m+1}\Big\} +$$

$$+ \Big\{(2\nabla_x + \nabla_x^{\,2} + \nabla_x^{\,3} + \dots)\bar{f}_i^{m+1} - 2h\,\bar{g}_i^{m+1}\Big\}\Big].$$

where it is understood that (4.2.10), (4.2.11) are to be used only at $i = N$. This replacement, by nature of its derivation, has the same principal part of truncation error as the method for the interior of the region. The highest power of $\nabla_x$ is determined as in the simpler equations.

The question of extrapolation or deferred approach to the limit once again arises naturally from the introduction of this boundary replacement, and with it the question of the existence of a second boundary technique which uses only alternate points.

This second boundary replacement can be determined along similar lines to that for the simpler equations. We do not derive the scheme here but point out some of the important points in its derivation. For $a = \frac{1}{4}$, that is the extension of the Lax-Wendroff method to the general quasilinear equations, the predictor formula is found to be

$$(4.2.12)\quad u_{m+1}^{*} = \tfrac{1}{2}\big[(1 + \tilde{a}_0)u_{i-1}^m + \tilde{a}_1 u_{i-3}^m + \tilde{a}_2 u_{i-5}^m\big] + \frac{p}{4}\big[(1 + a_0)f_{i-1}^m +$$

$$+ a_1 f_{i-3}^m + a_2 f_{i-5}^m + \dots + 2h\,g_i^m\big],$$

where $i = N$. The corrector, on substitution of $a = \frac{1}{4}$, is found to be

(4.2.13) $u_{m+1} = u_m - \frac{D}{4}[2H_x f_i^{*m+1} + H_x \bar{f}_i^{m+1} - H_x f_i^m - 2h(2\gamma_i^{*m+1} + \bar{\gamma}_i^{m+1} - \gamma_i^m)]$

which, as may easily be verified, is equivalent to

(4.2.14) $u_{m+1} = u_m - \frac{D}{2}[H_x \tilde{f}_i^{*m+1} - 2h\gamma_i^{*m+1}]$ ,

where

$$\tilde{f}_i^{*m+1} = f(u_i^{*m+1}, ih, (m + \tfrac{1}{2})k) \text{ and}$$

$$\gamma_i^{*m+1} = \gamma(u_i^{*m+1}, ih, (m - \tfrac{1}{2})k).$$

Now applying the alternate point boundary correction technique to the corrector in the form (4.2.14), we obtain

(4.2.15) $u_{m+1} = u_m + \frac{D}{2}[(1 + a_0)\tilde{f}_{i-1}^{*m+1} + a_2\tilde{f}_{i-5}^{*m+1} + \dots + 2h\gamma_i^{*m+1}]$

where $i = N$.

Similar boundary techniques may be derived for the system of equations (4.1.2). Their form is obvious and will not be quoted. Similarly, the form of the boundary techniques for problems with negative eigenvalues is easily seen and will not be described.

## 4.3 The one dimensional implicit method.

Having considered the explicit methods and associated boundary techniques for the solution of (4.1.1) and (4.1.2), we now proceed to derive implicit methods which are of predictor-corrector form and which are second order accurate.

Consider, for the system (4.1.1), the implicit general predictor-corrector method

$$(4.3.1)\quad u^*_{m+1} = \mu_x u_m - ap(H_x f_i^m - 2h\gamma_i^m)$$

$$(4.3.2)\quad u_{m+1} = u_m - p[bH_x f_i^m + cH_x \tilde{A}_i^{*m+1} u_i^{m+1} + dH_x \bar{f}_i^{m+1} +$$

$$+ 2h(\alpha\gamma_i^m + \beta\gamma_i^{*m+1} + \delta\bar{\gamma}_i^{m+1})] ,$$

where $f(u, x, t) \equiv \tilde{A}(u, x, t).u$. (We note that this relation does not uniquely define $\tilde{A}(u, x, t)$) and a, b, c, d, $\alpha$, $\beta$, $\delta$ are constants to be determined. Eliminating $u^*_{m+1}$ between (4.3.1) and (4.3.2), we get

$$(4.3.3)\quad u_{m+1} = u_m - p[bH_x f_i^m + cH_x(\tilde{A}(u_m - ap(H_x f_i^m - 2h\gamma_i^m))u_i^{m+1} + dH_x \bar{f}_i^{m+1} +$$

$$+ 2h(\alpha\gamma_i^m + \beta\gamma(u_m - ap(H_x f_i^m - 2h\gamma_i^m)) + \delta\bar{\gamma}_i^{m+1})] + O(h^3).$$

Expanding (4.3.3) further using Taylor's series, retaining terms up to and including those of order $h^2$, we obtain

$$u_{m+1} = u_m - p\Big[(b + c + d)H_x f - capH_x\Big\{H_x f - 2h\gamma\Big)\frac{\partial \tilde{A}}{\partial u}.u\Big\} + cphH_x\Big\{\tilde{A}\Big(\gamma - \frac{\partial f}{\partial x}\Big)\Big\}$$

$$+ dphH_x\frac{\partial f}{\partial t} + 2h(\alpha + \beta + \delta)\gamma - 2aph\frac{\partial \gamma}{\partial u}(H_x f - 2h\gamma) + 2ph^2\gamma\frac{\partial \gamma}{\partial t}\Big]_m + O(h$$

where (4.1.1) and $\tilde{A}.u = f$ have been used. Therefore

$$(4.3.4)\quad u_{m+1} = u_m - 2ph\Big[(b + c + d)\frac{\partial f}{\partial x} + (\alpha + \beta + \delta)\gamma\Big]_m$$

$$+ 4p^2h^2\Big[ca\Big\{\Big(\frac{\partial f}{\partial x} - \gamma\Big)\frac{\partial \tilde{A}}{\partial u}.u\Big\}_x - \tfrac{1}{2}c\Big\{\tilde{A}\Big(\gamma - \frac{\partial f}{\partial x}\Big)\Big\}_x - \tfrac{1}{2}d\frac{\partial}{\partial x}\frac{\partial f}{\partial t} +$$

$$+ a\Big(\frac{\partial f}{\partial x} - \gamma\Big)\frac{\partial \gamma}{\partial u} - \tfrac{1}{2}\gamma\frac{\partial \gamma}{\partial t}\Big]_m + O(h^3)$$

where the subscript x defines the first partial derivative with respect to x. If $u_{m+1}$ is now expanded in terms of $u_m$ and its derivatives using Taylor's series, then

$$(4.3.5)\quad u_{m+1} = u_m + ph\Big(\gamma - \frac{\partial f}{\partial x}\Big)_m + \frac{p^2h^2}{2}\Big[\frac{\partial \gamma}{\partial t} + \frac{\partial \gamma}{\partial u}\Big(\gamma - \frac{\partial f}{\partial x}\Big) - \frac{\partial}{\partial x}\frac{\partial f}{\partial t} -$$

$$- \frac{\partial}{\partial x}\Big\{\frac{\partial f}{\partial u}\Big(\gamma - \frac{\partial f}{\partial x}\Big)\Big\}\Big]_m + O(h^3),$$

where (4.1.1) has been used. If, in addition, we note that (formally)

$$\frac{\partial f}{\partial u} = \frac{\partial}{\partial u}(\tilde{A}.u) = \frac{\partial \tilde{A}}{\partial u}.u + \tilde{A},$$

then

$$(4.3.6)\quad u_{m+1} = u_m + ph\left(g - \frac{\partial f}{\partial x}\right)_m + \frac{p^2h^2}{2}\left[\frac{\partial g}{\partial t} + \frac{\partial g}{\partial u}\left(g - \frac{\partial f}{\partial x}\right) - \frac{\partial}{\partial x}\frac{\partial f}{\partial t} - \frac{\partial}{\partial x}\left(\frac{\partial A}{\partial u}.u\left(g - \frac{\partial f}{\partial x}\right) + A\left(g - \frac{\partial f}{\partial x}\right)\right)\right]_m + O(h^3).$$

Comparing coefficients between (4.3.4) and (4.3.6), it may be seen that in order that (4.3.1, 4.3.2) be a second order correct approximation to (4.1.1

$$-2(b + c + d) = -1$$
$$-2(\alpha + \beta + \gamma) = 1$$
$$4ca = \tfrac{1}{2}$$
$$-2c = -\tfrac{1}{2}$$
$$-2d = -\tfrac{1}{2}$$
$$4a\beta = -\tfrac{1}{2}$$
$$-2\gamma = \tfrac{1}{2}$$

must be satisfied. This set of simultaneous equations has the solution

$$-\beta = c = d = -\gamma = \frac{1}{4}, \quad \alpha = b = 0, \quad a = \tfrac{1}{2}.$$

The resulting scheme is therefore

$$u^*_{m+1} = \mu_x u_m - \frac{p}{2}(H_x f^m_i - 2h\, g^m_i)$$

$$u_{m+1} = u_m - \frac{p}{4}[H_x \bar{A}^{*m+1}_i\, u^{m+1}_i + H_x \bar{f}^{*m+1}_i - 2h(g^{*m+1}_i + \bar{g}^{m+1}_i)]$$

which may be written in the more familiar implicit form

$$u^*_{m+1} = \mu_x u_m - \frac{p}{2}(H_x f^m_i - 2h\,{}^m_i)$$

(4.3.7)

$$[I + \frac{p}{4}H_x A^*_{m+1}]u_{m+1} = u_m - \frac{p}{4}[H_x \bar{f}_{m+1} - 2h(\zeta^*_{m+1} + \bar{g}_{m+1})]$$

where I is the unit matrix. This implicit scheme requires the inversion of a block tridiagonal matrix which may be effected by any of the standard procedures; see, for example, Varga [68].

For the system of equations (4.1.2), a similar analysis yields an implicit predictor-corrector method of the form

$$u^*_{m+1} = {}_x u_m - \frac{p}{2}[A_m H_x u^m_i - 2h\,{}^m_i]$$

$$[I + \frac{p}{4}A^*_{m+1}H_x]u_{m+1} = u_m - \frac{p}{4}[\bar{A}_{m+1}H_x u^m_i - 2h({}^*_{m+1} + \bar{g}_{m+1})],$$

which again requires the inversion of a block tridiagonal matrix. We note, briefly, that the implicit methods, as for the simpler equations of Chapter 2, determine 'a' uniquely to be $\frac{1}{2}$. That is, the predictor level is always at the (m + 1)th time level.

## 4.4 Two dimensional explicit method.

Consider the two space dimensional system of quasilinear hyperbolic equations

$$(4.4.1)\quad \frac{\partial u}{\partial t} + \frac{\partial f}{\partial x}(u, x, y, t) + \frac{\partial g}{\partial y}(u, x, y, t) = \zeta(u, x, y, t),$$

where u is an n-component unknown vector function of x, y and t, and f, g and $\zeta$ are known functions of u, x, y and t. If we carry out the differentiation in (4.4.1) we obtain

$$(4.4.2)\quad \frac{\partial u}{\partial t} + A(u, x, y, t)\frac{\partial u}{\partial x} + B(u, x, y, t)\frac{\partial u}{\partial y} = \zeta(u, x, y, t),$$

(the $\zeta$'s in equations (4.4.1) and (4.4.2) not necessarily being the same)
where A and B are the Jacobian matrices of f and g, with respect to the components of u, respectively. A and B will be assumed to have a complete set of positive eigenvalues for simplicity, although the analysis may be carried through in general. Once again, whilst (4.4.2) may be derived from (4.4.1) the reverse is not always true and hence the methods for (4.4.2) will be given after the derivation of those for (4.4.1).

Consider the general predictor-corrector method

$$(4.4.3)\quad u^*_{m+1} = \tfrac{1}{2}(\mu_x + \mu_y)u_m - ap[H_x f_m + H_y g_m - 2h\zeta_m]$$

$$(4.4.4)\quad u_{m+1} = u_m - p[bH_x f_m + cH_x f^*_{m+1} + dH_x \bar{f}_{m+1} + eH_y g_m + fH_y g^*_{m+1} +$$

$$+ q\bar{g}_{m+1} + 2h(\alpha\zeta^*_{m+1} + \beta\zeta_m + \gamma\bar{\zeta}_{m+1})]$$

for the solution of (4.4.1). Carrying through the analysis in a similar manner to the derivation of previous schemes, the system of equations

$$
\begin{array}{llll}
2(b + c + d) = 1 , & & 2(e + \ell + q) = 1 \\
2(\alpha + \beta + \gamma) = -1 , & & 4ac = \frac{1}{2} \\
4a\alpha = -\frac{1}{2} & & 4a\ell = \frac{1}{2} \\
2d = \frac{1}{2} & & 2q = \frac{1}{2} \\
2\gamma = -\frac{1}{2} & & &
\end{array}
$$

is obtained, which must be satisfied if (4.4.3, 4.4.4) is to be second order accurate. This set of equations has the solution

$$b = e = -\beta = \frac{1}{4}(1 - \frac{1}{2a}), \qquad c = \ell = -\alpha = \frac{1}{8a} ,$$

$$d = q = -\gamma = \frac{1}{4} ,$$

so that the explicit method becomes

$$(4.4.5) \quad u^*_{m+1} = \tfrac{1}{2}(\mu_x + \mu_y)u_m - ap(H_x f_m + H_y g_m - 2h\gamma_m)$$

$$(4.4.6) \quad u_{m+1} = u_m - \frac{p}{4}[(1 - \frac{1}{2a})(H_x f_m + H_y g_m - 2h\gamma_m) + \frac{1}{2a}(H_x f^*_{m+1} + H_y g^*_{m+1} -$$

$$-2h\gamma^*_{m+1}) + (H_x \bar{f}_{m+1} + H_y \bar{g}_{m+1} - 2h\bar{\gamma}_{m+1})]$$

which is second order accurate. Similarly, the explicit scheme for (4.4.2) is

$$(4.4.7) \quad u^*_{m+1} = \tfrac{1}{2}(\mu_x + \mu_y)u_m - ap(A_m H_x u_m + B_m H_y u_m - 2h\gamma_m)$$

$$(4.4.8) \quad u_{m+1} = u_m - \frac{p}{4}[(1 - \frac{1}{2a})(A_m H_x u_m + B_m H_y u_m - 2h\gamma_m) + \frac{1}{2a}(A^*_{m+1} H_x u^*_{m+1} +$$

$$+ B^*_{m+1} H_y u^*_{m+1} - 2h\gamma^*_{m+1}) + (\bar{A}_{m+1} H_x u_m + \bar{B}_{m+1} H_y u_m - 2h\bar{\gamma}_{m+1})].$$

## 4.5 Generalizations of Boundary Techniques for two dimensional methods.

In section (4.2) we considered the generalization of the boundary techniques of Chapter 3 to the one dimensional analogues of (4.4.1) and (4.4.2). We will consider briefly the extensions to the methods for (4.4.1) and (4.4.2).

In the problem with positive eigenvalues, boundary data is required on $x = 0, \alpha$, $(0 \le y \le \beta)$ and $y = 0, \beta$, $(0 \le x \le \alpha)$ where the region of computation is defined to be $0 \le x \le \alpha$, $0 \le y \le \beta$. Initial data is given on the plane $0 \le x \le \alpha$, $0 \le y \le \beta$, $t = 0$, and on the planes $x = 0$, $y = 0$, $t > 0$. Hence we require the extra data on the planes $x = \alpha$ and $y = \beta$. The method of incorporating this extra data, as was seen earlier, was either

(1) Introduction of theoretical data and iteration of the corrector

or

(2) A boundary replacement consistent with the difference method used in the interior.

For method (1) the iteration method is

$$(4.5.1)\qquad \begin{aligned} u^*_{m+1} &= \tfrac{1}{2}(\mu_x + \mu_y)u_m - \tfrac{p}{2}[H_x f_m + H_y g_m - 2h\phi_m] \\ u^{(j'+1)}_{m+1} &= u_m - \tfrac{p}{4}[(H_x f^{(j')}_{m+1} + H_y g^{(j')}_{m+1} - 2h\phi^{(j')}_{m+1}) + (H_x \bar{f}_{m+1} + \\ &\qquad + H_y \bar{g}_{m+1} - 2h\bar{\phi}_{m+1})] \end{aligned}$$

for the solution of (4.4.1), and

$$u^{*}_{m+1} = \tfrac{1}{2}(\mu_x + \mu_y)u_m - \frac{D}{2}[A_m H_x u_m + B_m H_y u_m - 2h\gamma_m]$$

(4.5.2)

$$u^{(j'+1)}_{m+1} = u_m - \frac{D}{4}[(A^{(j')}_{m+1} H_x u^{(j')}_{m+1} + B^{(j')}_{m+1} H_y u^{(j')}_{m+1} - 2h\gamma^{(j')}_{m+1}) +$$

$$+ (\bar{A}_{m+1} H_x u_m + \bar{B}_{m+1} H_y u_m - 2h\bar{\gamma}_{m+1})]$$

for the solution of (4.4.2) where $j' = 1, 2, \ldots$ and $u^{(1)}_{m+1} \equiv u^{*}_{m+1}$. These methods (4.5.1) and (4.5.2) were derived by a similar analysis to the iteration method of section (4.2) and $a = \frac{1}{2}$ has been substituted.

For method (2) there are, once again, two separate boundary schemes available, one for $a = \frac{1}{4}$ and the other for other values of 'a'. For the latter case the relationships between the difference operators are used to "invert" (in the sense of section 3.3) $H_x$ and $H_y$. The schemes are derived as before and as in Chapter 2 for the simpler equations, there are three applications to consider:

(i) $x = \alpha$ $\quad 0 < y < \beta$

(ii) $y = \beta$ $\quad 0 < x < \alpha$

(iii) $x = \alpha,\ y = \beta$.

The schemes for (i), (ii) and (iii) are similar to 1), 2), 3) of section (3.3) and will not be quoted.

For $a = \frac{1}{4}$ in these two dimensional explicit methods, we note that the corrector may be written as

$$u_{m+1} = u_m - \frac{D}{2}[H_x \tilde{f}^{*m+1}_{ij} + H_y \tilde{g}^{*m+1}_{ij} - 2h\tilde{\gamma}^{*m+1}_{ij}] ,$$

where, as in section (4.2), $\tilde{f}^{*m+1}_{ij} = f(u^{*m+1}_{ij}, ih, jh, (m + \frac{1}{2})k)$ and similarly for $\tilde{g}^{*m+1}_{ij}$ and $\tilde{\gamma}^{*m+1}_{ij}$, so that the alternate-point boundary scheme is applied to

$$u^*_{m+1} = \tfrac{1}{2}(\mu_x + \mu_y)u_m - \tfrac{p}{4}[H_x f_m + H_y g_m - 2h\gamma_m]$$

$$u_{m+1} = u_m - \tfrac{p}{2}[H_x \tilde{f}^{*m+1}_{ij} + H_y \tilde{g}^{*m+1}_{ij} - 2h\tilde{\gamma}^{*m+1}_{ij}]$$

The required boundary scheme depends upon the lemma and theorem of Chapter 3; a similar result obviously holds for the y-direction, and its form is a natural extension of (4.2.12) and (4.2.15) where, once again, three variations of the scheme for (i), (ii) and (iii) must be considered. We stress that this alternate point boundary procedure need only be used if deferred approach to the limit is being considered, otherwise the adjacent point boundary technique may be used for all values of 'a'. If deferred approach to the limit is being considered, a process of determining the form of the asymptotic expansion may be carried out along similar lines to (3.8).

## 4.6 The two dimensional implicit method - the A.D.I. method.

We now consider the natural extension of the Alternating Direction method, considered in Chapter 2, for the solution of (4.4.1) and (4.4.2).

Consider first the system (4.4.1) and define

$$\tilde{A}(u) . u = f(u), \qquad \tilde{B}(u).u = g(u).$$

Consider the predictor-corrector method

$$(4.6.1) \quad u^{**}_{m+1} = \tfrac{1}{2}(\lambda_x + \lambda_y)u_m - ap[H_x f_m + H_y g_m - 2h\gamma_m]$$

$$(4.6.2) \quad u^{*}_{m+1} = u_m - bp[H_x f_m + H_y g_m] - cp[H_y \tilde{B}^{**}_{m+1} u^{*}_{m+1}] - ep[H_x \bar{f}_{m+1} + H_y \bar{g}_{m+1}]$$
$$+ dp^2 H_y \tilde{B}_m H_x f_m + 2ph(\alpha\gamma^{**}_{m+1} + \beta\bar{\gamma}_{m+1} + \delta\gamma_m)$$

$$(4.6.3) \quad u_{m+1} = u^{*}_{m+1} - \theta p H_x \tilde{A}^{**}_{m+1} u_{m+1},$$

where $\tilde{A}^{**}_{m+1} \equiv \tilde{A}(u^{**}_{m+1})$ and $\tilde{B}^{**}_{m+1} \equiv \tilde{B}(u^{**}_{m+1})$. The symbol ** denotes the predictor level (4.6.1) and * denotes an intermediate level of the overall corrector formula (4.6.2), (4.6.3). The quantities a, b, c, d, e, $\alpha$, $\beta$, $\delta$, $\theta$ are constants to be determined.

By eliminating $u^{**}_{m+1}$ between (4.6.1), (4.6.2) and (4.6.3) and then eliminating $u^{*}_{m+1}$ from the resulting equations, expanding by Taylor's series and finally comparing the result with the Taylor expansion of $u_{m+1}$, we obtain the following equations.

$$2(b + e + \theta) = 1 \ , \qquad 2(b + e + c) = 1$$

$$2(\alpha + \beta + \gamma) = 1 \ , \qquad 2e = \tfrac{1}{2}$$

$$4a\theta = \tfrac{1}{2} \ , \qquad 2\theta = \tfrac{1}{2}$$

$$4ac = \tfrac{1}{2} \ , \qquad 2c = \tfrac{1}{2}$$

$$4a\alpha = \tfrac{1}{2} \qquad 2\beta = \tfrac{1}{2}$$

$$-4c\theta + 4d = 0,$$

in order that (4.6.1), (4.6.2) and (4.6.3) be overall correct to order $h^2$. These equations have the solution

$$c = e = \theta = \alpha = \beta = \frac{1}{4}, \quad a = \tfrac{1}{2}, \quad \gamma = b = 0, \quad d = \frac{1}{16}.$$

The resulting method is therefore, -

$$u^{**}_{m+1} = \tfrac{1}{2}(\rho_x + \rho_y)u_m - \frac{p}{2}[H_x f_m + H_y g_m - 2h\gamma_m]$$

$$u^{*}_{m+1} = u_m - \frac{p}{4}[H_x \bar{f}_{m+1} + H_y \bar{g}_{m+1} + H_y \tilde{B}^{**}_{m+1} u^{*}_{m+1} - 2h(\gamma^{**}_{m+1} + \bar{\gamma}_{m+1})] +$$

$$+ \frac{p^2}{16} H_y \tilde{B}_m H_x f_m$$

$$u_{m+1} = u^{*}_{m+1} - \frac{p}{4} \tilde{A}^{**}_{m+1} u_{m+1},$$

which may be written in the D'Yakonov split form [12],

$$u^{**}_{m+1} = \tfrac{1}{2}(\rho_x + \rho_y)u_m - \frac{p}{2}[H_x f_m + H_y g_m - 2h\gamma_m]$$

$$(4.6.4) \quad [I + \frac{p}{4} H_y \tilde{B}^{**}_{m+1}] u^{*}_{m+1} = u_m - \frac{p}{4}[H_x \bar{f}_{m+1} + H_y \bar{g}_{m+1} - 2h(\gamma^{**}_{m+1} + \bar{\gamma}_{m+1})] +$$

$$+ \frac{p^2}{16} H_y \tilde{B}_m H_x f_m$$

$$[I + \frac{p}{4} H_x \tilde{A}^{**}_{m+1}]u_{m+1} = u^*_{m+1}.$$

The corresponding alternating direction method for the system of equations (4.4.2) is derived in a similar manner. It is found to be

$$u^{**}_{m+1} = \frac{1}{2}(\mu_x + \mu_y)u_m - \frac{p}{2}[A_m H_x u_m + B_m H_y u_m - 2h\gamma_m]$$

$$[I + \frac{p}{4} B^{**}_{m+1} H_y]u^*_{m+1} = u_m - \frac{p}{4}[\bar{A}_{m+1} H_x u_m + \bar{B}_{m+1} H_y u_m -$$

$$- 2h(\gamma^{**}_{m+1} + \bar{\gamma}_{m+1})] + \frac{p^2}{16} B_m H_y A_m H_x u_m$$

$$[I + \frac{p}{4} A^{**}_{m+1} H_x]u_{m+1} = u^*_{m+1}.$$

## 4.7 The explicit alternating direction method.

As in the simpler equations

$$\frac{\partial u}{\partial t} + \frac{\partial f}{\partial x} + \frac{\partial g}{\partial y} = 0,$$

and its non-conservation counterpart, an explicit alternating direction method for the systems of equations (4.4.1) and (4.4.2) may be derived. These methods will be the non-linear analogues of the methods considered by Gourlay and Mitchell [23]. It was found in section (2.6) that considerable care had to be taken in evaluating the arguments of the matrices. With the introduction of the inhomogeneous term $\gamma(u, x, y, t)$, the care in evaluating arguments must be extended to $\gamma$. However, since difference operators do not act on $\gamma$, there is considerably less difficulty in choosing the average of the argument. The methods are derived along similar lines to past methods in this thesis and so the derivation will be omitted. The scheme for (4.4.1) is found to be

$$(4.7.1) \quad u^{**}_{m+1} = \mu_x \mu_y u_m - \frac{p}{2}[\mu_y H_x f^m_{ij} + \mu_x H_y g^m_{ij} - 2h\gamma(\mu_x \mu_y u^m_{ij})]$$

$$(4.7.2) \quad [\mu_y + \frac{p}{4} H_y \tilde{B}(\mu_y u^{**m+1}_{ij})] u^{*m+1}_{ij} = [\mu_y - \frac{p}{4} H_y \tilde{B}(\mu_x u^m_{ij})][\mu_x - \frac{p}{4} H_x \tilde{A}(u^m_{ij})] u^m_{ij} +$$

$$+ \frac{p}{2}[\gamma^{**}_{m+1} + \bar{\gamma}_{m+1}(\mu_x \mu_y u^m_{ij})]$$

$$(4.7.3) \quad [\mu_x + \frac{p}{4} H_x \tilde{A}(\mu_x \mu_y u^{**m+1}_{ij})] u^{m+1}_{ij} = u^{*m+1}_{ij} ,$$

and for (4.4.2) to be

$$(4.7.4)\quad u^{**}_{m+1} = {}_x{}_y u_m - \frac{p}{2}[\;{}_y A({}_x u^m_{ij})H_x u^m_{ij} + {}_x B({}_y u^m_{ij})H_y u^m_{ij}$$

$$- 2h\,({}_x{}_y u^m_{ij})]$$

$$(4.7.5)\quad [\;{}_y + \frac{p}{4}B(u^{**m+1}_{ij})H_y]u^{*m+1}_{ij} = [\;{}_y - \frac{p}{4}B({}_x{}_y u^m_{ij})H_y][\;{}_x - \frac{p}{4}A({}_x u^m_{ij})H_x]u^m_{i}$$

$$+ \frac{p}{2}[\;{}^{**}_{m+1} + \bar{\;}_{m+1}({}_x{}_y u^m_{ij})]$$

$$(4.7.6)\quad [\;{}_x - \frac{p}{4}A(\mu_y u^{**m+1}_{ij})H_x]u^{m+1}_{ij} = u^{*m+1}_{ij}$$

The schemes ((4.7.1), (4.7.2), (4.7.3)), ((4.7.3), (4.7.4), (4.7.5)) are explicit in nature and involve four points at each time level. Once again (4.7.3) and (4.7.6) must be noted. The range of i and j do not correspond to the ranges of i and j in the first two formulae of each method. It is for this reason that care must be taken in evaluating the arguments of the matrices - to make sure that the argument u <u>has</u> actually been calculated, and it does not lie outside the grid of mesh size 2h.

For the arguments of the matrices in the ** level, in particular $\tilde{B}$ in (4.7.2) and $\tilde{A}$ in (4.7.3), values of u** are required outside the region of computation. These may be evaluated using any of the boundary techniques mentioned in this thesis. They will not be expounded here as they may be performed in an obvious way.

## 4.8 Stability of the one and two dimensional methods.

In extending the finite difference methods derived in Chapter 2, we have had to consider several factors. Firstly, the vectors f and g have now been allowed to be functions of t in addition to u, x and y. In order to determine the stability characteristics of the methods derived in Chapter 2 we had to linearize the methods. We see therefore, that when linearizing the present schemes, the introduction of x, y and t into f and g does not affect the linearized stability analysis. Second, we have allowed lower order terms $\gamma$ into the system of equations. Since the stability is a limiting process as h → o, it can be seen from the difference methods, that the vector functions of $\gamma$ are premultiplied by h so that these terms tend to zero as h → o, and thus the vector functions of $\gamma$ have no effect on the stability. The linearized stability characteristics of the difference schemes of sections (4.1, 4.3, 4.4, 4.6, 4.7) are exactly those of the corresponding methods of Chapter 2. Thus the stability conditions for these methods are as follows:

(i) one dimensional explicit: stable for $a \geq \frac{1}{4}$, $p|\lambda| \leq \frac{1}{\sqrt{a}}$ where $|\lambda|$ is the maximum modulus eigenvalue of the matrix A.

(ii) one dimensional implicit: unconditionally stable in linearized sense

(iii) two dimensional explicit: conditionally stable, the condition being obtained only for specific problems.

(iv) A.D.I. method: unconditionally stable in linearized sense.

(v) explicit A.D. method: unconditionally stable in linearized sense (by direct extension of Gourlay and Mitchell [23]).

Because the introduction of the new terms into the differential system does not affect the linearized analysis we see that the convergence analysis (which is again a linearized one) of the iteration of the corrector will be analogous to that given in Chapter 2, and similar conditions will be derived.

## 4.9 Numerical Results

To test the methods discussed in this chapter, we used problems for which the theoretical solutions were known and compared the results obtained by the numerical methods with the theoretical solutions. To test the one dimensional methods of sections (4.1) and (4.3), we considered the equation

$$\frac{\partial u}{\partial t} + \frac{x^2 u^2}{t} \frac{\partial u}{\partial x} = \left(\frac{2x^3 u^2}{t} - 1\right) \cos\,(x^2 - t),$$

subject to the initial condition $u(x,\ 1) = \sin\,(x^2 - 1)$ and the boundary condition $u(0,\ t) = -\sin t$. This problem has the theoretical solution

$$u(x,\ t) = \sin\,(x^2 - t),$$

and was solved in the region $o \leq x \leq 1,\ 1 \leq t \leq 1 + 100k$, where $k = ph$. Therefore the space increment $h$ was chosen to be 0.1 and $p$ took the values 0.6 and 1.0. The problem was solved in four ways.

First, the explicit general predictor-corrector method was used with $a = \frac{1}{4}$ and $\frac{1}{2}$, with theoretical data substituted at $x = 1$ for the predictor and the corrector levels.

Second, the same method was used with $a = \frac{1}{4}$ and $\frac{1}{2}$ but with the boundary data given by the technique of (4.2) with $\gamma = 2$ at the predictor level and $\gamma = 3$ at the corrector level for $x = 1$.

Third, the iterative method was used when the corrector was iterated twice.

Finally the one dimensional implicit method was used. The differences between the computed and theoretical solutions are given in table (4.9.1) and are quoted at $x = 0.7$.

To test the two dimensional methods of (4.4) and (4.6) we considered the equation

$$\frac{\partial u}{\partial t} + ue^{-t}(1 + x + x^2 - y)\frac{\partial u}{\partial x} + ue^{-t}(1 + y + y^2 - x)\frac{\partial u}{\partial y} =$$

$$= - u(1 + e^{-2t}(2 + x^2 + y^2)) \text{ ,}$$

subject to the conditions

$$u(x, y, o) = 1 - x - y, \quad u(o, y, t) = (1 - y)e^{-t},$$
$$u(x, o, t) = (1 - x)e^{-t}$$

which has the theoretical solution $u = (1 - x - y)e^{-t}$. This problem was solved in the region $o \leq x, y \leq 1 \quad o \leq t \leq 100k$ using the explicit general predictor-corrector method and the A.D.I. method. As in the one dimensional problem, $p$ took the values 0.6 and 1.0 and 'a' took the values $\frac{1}{4}$ and $\frac{1}{2}$ in the explicit method. The mesh size h was chosen to be 0.1. The problem was solved using the same class of boundary procedures as the one dimensional problem, that is, theoretical data at predictor and corrector levels for $x = 1$, $y = 1$; boundary replacement of appropriate form at the predictor and corrector levels; iteration of the corrector scheme. The errors are quoted in table (4.9.2).

We next considered the applications of the boundary techniques of (3.3) and (3.5) together with the methods of this chapter to the deferred approach to the limit discussed in (3.8) for the problem

$$(4.9.1)\quad \frac{\partial v}{\partial t} + \frac{1}{2}\frac{\partial}{\partial x}\left(v + \frac{1}{1+t}\right)^2 = \frac{1}{(1+t)^2} \qquad \begin{cases} v(x, o) = x, & o < x < 1 \\ v(o, t) = o, & t > o \end{cases}$$

which has the theoretical solution

$$v = \frac{x}{1 + t}.$$

The value of $\gamma$ was taken to be 6. The results of using the boundary procedure of (3.3) are given in Table (4.9.3), where just two extrapolations are given. Finally for the same problem using the boundary scheme of (3.5) the errors before and after extrapolation are shown in table (4.9.4).

Tables of Results

One dimensional problem: table (4.9.1)

| | | Explicit methods | | | Implicit methods |
|---|---|---|---|---|---|
| a | p | THEORETICAL BOUNDARY | TWO ITERATIONS | BOUNDARY TECHNIQUE | THEORETICAL BOUNDARY |
| 0.25 | 0.6 | $.1277 \times 10^{-1}$ | - | $.1997 \times 10^{-3}$ | - |
| | 1.0 | $.1623 \times 10^{-1}$ | - | $.1187 \times 10^{-2}$ | - |
| 0.5 | 0.6 | $.7222 \times 10^{-2}$ | $.4131 \times 10^{-2}$ | $.3236 \times 10^{-2}$ | $.2231 \times 10^{-3}$ |
| | 1.0 | $.5367 \times 10^{-2}$ | $.2339 \times 10^{-2}$ | $.2559 \times 10^{-3}$ | $.1629 \times 10^{-2}$ |

<u>Two dimensional problem: (table (4.9.2)</u>

| | | Explicit methods | | | Implicit methods |
|---|---|---|---|---|---|
| a | p | Theoretical Boundary | Two Iterations | Boundary Technique | A.D.I. |
| 0.25 | 0.6 | * | - | $-.4151 \times 10^{-4}$ | - |
| | 1.0 | * | - | $-.1102 \times 10^{-1}$ | - |
| 0.5 | 0.6 | * | $-.1635 \times 10^{-5}$ | $-.1175 \times 10^{-5}$ | $-.3290 \times 10^{-5}$ |
| | 1.0 | * | $-.5322 \times 10^{-6}$ | $-.4302 \times 10^{-6}$ | $-.1120 \times 10^{-6}$ |

<u>Table (4.9.3)</u>. Problem (4.9.1) solved by the general predictor-corrector method of this ~~next~~ chapter and using the boundary technique of (3.3).

ERRORS at $100 \times p \times h_0$ at $x = \frac{1}{2}$

| a | p | $h_0 = 0.05$ | $h_1 = 0.025$ | $h_3 = 0.0125$ | 2 Extrapolations |
|---|---|---|---|---|---|
| | 0.3 | $.16 \times 10^{-4}$ | $.17 \times 10^{-5}$ | $.50 \times 10^{-6}$ | $.21 \times 10^{-6}$ |
| 0.25 | 0.6 | $.13 \times 10^{-3}$ | $.14 \times 10^{-4}$ | $.79 \times 10^{-6}$ | $.11 \times 10^{-5}$ |
| | 1.0 | $.47 \times 10^{-3}$ | $.61 \times 10^{-4}$ | $.34 \times 10^{-5}$ | $.75 \times 10^{-5}$ |
| | 0.3 | $.25 \times 10^{-5}$ | $.61 \times 10^{-6}$ | $.15 \times 10^{-6}$ | $.54 \times 10^{-9}$ |
| 0.5 | 0.6 | $.17 \times 10^{-5}$ | $.45 \times 10^{-6}$ | $.11 \times 10^{-6}$ | $.54 \times 10^{-8}$ |
| | 1.0 | $.14 \times 10^{-6}$ | $.30 \times 10^{-6}$ | $.82 \times 10^{-7}$ | $.39 \times 10^{-8}$ |

Table (4.9.4): As for table (4.9.3) except using boundary technique of (3.5)

ERRORS at $100 \times p \times h_0$ at $x = \frac{1}{2}$

| a | p | $h_0 = 0.05$ | $h_1 = 0.025$ | $h_3 = 0.0125$ | 2 Extrapolations |
|---|---|---|---|---|---|
| | 0.3 | $.25 \times 10^{-5}$ | $.63 \times 10^{-6}$ | $.16 \times 10^{-6}$ | $.21 \times 10^{-8}$ |
| 0.25 | 0.6 | $.18 \times 10^{-5}$ | $.47 \times 10^{-6}$ | $.12 \times 10^{-6}$ | $.37 \times 10^{-9}$ |
| | 1.0 | $.13 \times 10^{-5}$ | $.34 \times 10^{-6}$ | $.84 \times 10^{-7}$ | $.11 \times 10^{-9}$ |
| | 0.3 | $.25 \times 10^{-5}$ | $.61 \times 10^{-6}$ | $.15 \times 10^{-6}$ | $.22 \times 10^{-9}$ |
| 0.5 | 0.6 | $.17 \times 10^{-5}$ | $.45 \times 10^{-6}$ | $.11 \times 10^{-6}$ | $.86 \times 10^{-8}$ |
| | 1.0 | $.76 \times 10^{-5}$ | $.30 \times 10^{-6}$ | $.82 \times 10^{-7}$ | $.31 \times 10^{-8}$ |

## 4.10 Comments on the methods and results of Chapter 4.

The introduction of the extra theoretical boundary data in the explicit methods introduces errors which, if they do not swamp the solution through non-linear instability, certainly prove detrimental. It is for this reason that we conclude that the boundary procedures, either the boundary replacement or iteration of the corrector have much to offer in that the error distribution over a line (or plane) at a given time level tends to be "smoother" than if theoretical boundary data were used. Of the two, we prefer the boundary replacement methods because they prove to be considerably less time consuming than the iterative method.

We now point out a fact, which although obvious is nevertheless important, that if the differential system in either one or two dimensions reduces to a simpler system with time independent coefficient matrices, then the difference methods introduced likewise reduce to the corresponding difference algorithms for the simpler equations. For example, if the system of equations is given by

$$\frac{\partial u}{\partial t} + A(u, x) \frac{\partial u}{\partial x} = b(u, x) \quad , \tag{4.10.1}$$

then the explicit difference scheme is obtained as a special case of the general one dimensional explicit predictor-corrector method of this chapter when

$$\bar{A}_{m+1} \longrightarrow A_m$$

and

$$\bar{\psi}_{m+1} \rightarrow \psi_m,$$

that is

$$u^*_{m+1} = \mu_x u_m - ap[A_m H_x u_m - 2h\psi_m]$$

$$u_{m+1} = u_m - \frac{p}{4}[(2 - \frac{1}{2a})(A_m H_x u_m - 2h\psi_m) + (A^*_{m+1} H_x u^*_{m+1} - 2h\psi^*_{m+1})]$$

and similarly for other problems.

The methods given in this chapter may be extended in a natural way to a higher number of space dimensions.

# CHAPTER 5

# PROBLEMS WITH DISCONTINUITIES: EXTENSIONS

GOURLAY AND MORRIS [26]

## Introduction.

So far, we have considered the above finite difference methods applied only to problems which have had continuous initial data and smooth solutions. As stated earlier (Chapter (1)) it is well known that owing to the non-linearity of the problems, no smooth solution will in general exist for all time (Lax [41]), that is, at some point a jump discontinuity (shock) may arise. Because the finite difference methods approximate the weak solutions defined in section (1.3), which in turn satisfy the Rankine-Hugoniot relations (for example, see Courant and Friedrichs [97]) across the shock, (Lax [42],), no special treatment is required to handle the shock and computation continues virtually as though there were no shock present. This, in itself, of course does not say that the features or the position of the shock are given with any great accuracy. We therefore require methods which represent the shock as accurately (featurewise and positionwise) as possible in that the shock is confined to a narrow band of 2-3 mesh points.

This discussion also applies to so called "weak discontinuities", that is, to problems which may have discontinuities in the derivatives of their solutions. These discontinuities may arise in several ways, for example, a Lipschitz discontinuity in the initial data or a weak discontinuity arising at some time out of analytic initial data as shown by Jeffrey [33]. These weak discontinuities may in turn give rise to the jump discontinuities discussed above.

It is well known (Lax [41]) that weak discontinuities arising in problems propagate along characteristics passing through the point of evolution of the discontinuity until such time that the weak discontinuity produces a jump discontinuity, the propagation of which is entirely different to that of the weak discontinuity. For the weak discontinuity problems, the method of characteristics would appear to be admirable for the solution of these problems up to the point of origin of the jump discontinuity. In fact, in several papers [33], [34], [35] Jeffrey has used this method to determine the time and distance to the formation of jump discontinuities. Thereafter one requires the exact position of the shock in order to compute beyond the shock. (This of course can be done using the above mentioned results of Jeffrey but one still cannot handle shocks automatically as in the case of finite difference methods.)

Because, in general, characteristics are curvilinear, finite difference methods applied on a rectangular grid will not represent exactly a discontinuity which propagates along a characteristic. A finite difference method can be considered 'good' in representing a discontinuity if it confines the discontinuity to a narrow band of 2-3 mesh points (Lax [42]). We therefore consider briefly in this final chapter the solution of problems with either of the discontinuities mentioned above present, in an effort to determine the effectiveness of the methods introduced here to represent shocks. This is done by comparing the methods with the two step Lax-Wendroff scheme (which is, of course, a special case of the methods) the ability of which to represent shocks is well-known, for example, see Gary [18], Burstein [5], Thommen [66], [67].

The idea of 'pseudo-viscosity' has purposely been omitted. This method of treating shocks was first introduced by Von Neumann and Richtmyer [69]. We do so because we require to determine the character of the methods without additional features, which may be added in a number of ways.

## 5.1 Discontinuity in the dependent variable.

The problem chosen was a simple one dimensional problem which had a discontinuity in the dependent variable u in the initial data such that the dependent variable $u(x, o) = 1$ for $o \leq x \leq \alpha$ and $u(x, o) = o$ for $\alpha < x \leq Nh$ where we took $N (= 50)$ grid points at each time level. This initial data was for the first order one dimensional conservation law

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}(\tfrac{1}{2}u^2) = 0.$$

The problem is in fact equivalent to that considered by Gary [18] formulated as

$$u(x, o) = \begin{cases} u_o, & x_1 \leq x \leq \alpha \\ q(x) = u_o + \dfrac{x - \alpha}{\beta - \alpha}(u_1 - u_o) & \\ u_1, & \beta \leq x \leq x_m \end{cases}$$

where we have chosen $x_1 = 0$ and $u_o = 1$, $q(x) = q(\beta) = u_1$, where $u_1 = 0$, $x_m = Nh$ and $u(o, t) = u_o = 1$ for all time. $\beta$ was chosen to be 0.1. The discontinuity occurring in the initial data is propagated into the field of computation with increasing time. We require to discover whether the one dimensional methods described in Chapter 2 confine the shock to 2-3 mesh lengths.

The results of the computation with the general predictor-corrector method

(5.1.1) $u^*_{m+1} = \mu_x u_m - apH_x f_m$

(5.1.2) $u_{m+1} = u_m - \frac{p}{2}[(1 - \frac{1}{4a})H_x f_m + \frac{1}{4a} H_x f^*_{m+1}]$

for $a = \frac{1}{4}, \frac{1}{2}$ and 1 are shown in graphs 1, 2, 3 for $p = 1.0$ and $h = 0.01$. As a comparison, the one dimensional implicit predictor-corrector scheme

(5.1.3) $u^*_{m+1} = \mu_x u_m - \frac{p}{2} H_x f_m$

(5.1.4) $[1 + \frac{p}{4} H_x \tilde{A}^*_{m+1}] u_{m+1} = u_m - \frac{p}{4} H_x f_m$

where $\tilde{A} = u/2$ is shown in graph 4. The graphs show the values of u at 10 time steps and 50 time steps. The two broken line graphs in the first figure are those drawn through alternate points given by the two step Lax-Wendroff method ($a = \frac{1}{4}$). This was done as it was realised (Richtmyer [58]) that because of the uncoupling between adjacent points, the Lax-Wendroff method yields two solutions in this problem; one being the 'image' of the other. As Richtmyer points out, in normal circumstances one can leave out one set of points. However in this example, which set of results does one leave out? If one leaves out the "earlier" graph (that is the one appearing to the left) then the position of the shock is given by the 'later' one and vice-versa. This must give some degree of doubt as to the position of the shock. If we consider the position of the shock obtained from $a = \frac{1}{2}$, it may be seen that this position compares well with an <u>average graph</u> of the two broken line graphs of $a = \frac{1}{4}$. If one calculates the position of the shock at

50 time steps, remembering that there is a 2h grid in operation it may easily be seen that the discontinuity is at $x = 0.35 \doteq 35 \times h$. Considering the positions given in the graphs 1, 2, 3, 4 it may be seen that for $a = 1$ the discontinuity lies between 34 and 37 steps, for $a = \frac{1}{2}$ between 35 and 37 steps, for $a = \frac{1}{4}$ between 34 and 38 steps (the average of the two graphs being similar to $a = \frac{1}{2}$) and the implicit between 34 and 36 steps.

Apart from the 'image' effect of the Lax-Wendroff method there appears, from the results quoted here, to be little to choose between the methods for the shock representation. The four methods suffer from the oscillations behind the shock. This phenomenon has been well known for the Lax-Wendroff method [54] and has been found to be eliminated to a great extent by the use of 'pseudo-viscosity'. Similar graphs were obtained for smaller values of p; these are not quoted. The explicit methods were used in conjunction with the boundary replacement technique of section (3.3).

## 5.2 Discontinuity in the derivative $u_x$ .

The second problem we considered was the conservation law

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{u^2}{2}\right) = 0$$

subject to the initial conditions

$$u(x, o) = \begin{cases} \alpha - x & \text{for } o \leq x \leq \alpha \\ o & \text{for } \alpha \leq x \leq Nh \end{cases}$$

using methods (5.1.1), (5.1.2) and (5.1.3) where N was again taken to be 50 and h = 0.1. The problem has continuous initial data but discontinuous $u_x$ at $x = \alpha$ where $\alpha$ was again chosen to be 0.1.

It may be easily verified that the characteristics of this problem for $o \leq x \leq \alpha$ are given by $\frac{dx}{dt} = \frac{\alpha - x}{1 - t}$ . The discontinuity will propagate along the characteristic issuing from the point of discontinuity. In the discrete case the discontinuity must lie between $x = \alpha$ and $x = \alpha - h$ and hence the discontinuity will propagate along a characteristic issuing between these points.

The solution, as may be easily verified, is not defined for $t \geq 1$. The discontinuity in the first derivative is stationary since the characteristic passing through $x = \alpha$ is parallel to the x-axis. However, after a shock forms at $t = 1$, the propagation of the shock is no longer governed by the characteristics and the shock will move. This is borne out by graph 5. In the graph the dependent variable u has been drawn for $a = \frac{1}{4}, \frac{1}{2}, 1$ and the implicit method at 5 and 50 time steps. The discontinuity is confined to 2-3 mesh lengths although the oscillations behind the shock are still present.

The results shown are for the largest mesh ratio taken, namely p = 1.0. The experiments were also carried out for smaller mesh ratios when similar graphs were obtained, the discontinuity being confined to two or three mesh lengths but the oscillations were again present behind the front. We were not unduly worried about these oscillations, as experiments by other authors (see references above) using the Lax-Wendroff method together with various pseudo-viscous terms have proved so successful that we feel that similar terms added to the other methods mentioned here will yield considerable smoothing of the oscillations, whilst retaining the narrow region of the shock. All of the explicit methods used the boundary technique of section (3.3).

## 5.3 Concluding remarks and extensions.

The finite difference methods introduced here can be extended to M space dimensions with considerable ease in an obvious manner. For example the explicit method for the system of conservation laws in M space dimensions

$$(5.3.1) \qquad \frac{\partial u}{\partial t} + \sum_{j=1}^{M} \frac{\partial f_j}{\partial x_j} = 0$$

would obviously be

$$u^*_{m+1} = \frac{1}{M} \sum_{j=1}^{M} \mu_{x_j} u_m - ap \left[ \sum_{j=1}^{M} H_{x_j} f^m_j \right]$$

$$u_{m+1} = u_m - \frac{p}{2}\left[\left(1 - \frac{1}{4a}\right) \sum_{j=1}^{M} H_{x_j} f^m_j + \frac{1}{4a} \sum_{j=1}^{M} H_{x_j} f^{*m+1}_j\right],$$

where $\mu_{x_j}$ is the usual averaging difference operator in the $x_j$ direction, and we have suppressed the spacial indices on the vectors so as not to cause confusion (the $f_j$ here must not be confused with the $f_i$ of the methods in earlier chapters - here it refers to the j'th function).

Similarly the D'Yakonov splitting introduced for the non-linear problems will extend in an obvious manner for (5.3.1) using the two dimensional method as a guide. The extensions of the explicit and implicit methods carry over also for the general equations

$$\frac{\partial u}{\partial t} + \sum_{j=1}^{N} \frac{\partial f_j}{\partial x_j} = \zeta(u, x_1, x_2, \ldots, x_n, t)$$

and their quasilinear counterparts.

By way of a final generalization we consider the extension, considered by Strang [62] of the one dimensional Lax-Wendroff method, to a two dimensional version which is 'optimum' in that the stability condition (which in Chapter 2 was equivalent to $p|\lambda_A|, p|\lambda_B| = \frac{1}{2\sqrt{2}}$) is eased to the Courant-Friedrichs-Lewy condition $p|\lambda_A|, p|\lambda_B| \leq 1$. The condition for stability of Chapter 2 may prove to be too severe in many practical cases and it is of the purpose of this last consideration to derive a scheme which is optimum.

In order to alleviate such severe stability restrictions, several authors have considered modifications of the two step Lax-Wendroff method, which when written on a grid of mesh size h, is of the form

$$u^*_{m+1} = \tfrac{1}{2}(\mu_x + \mu_y)u_m - \tfrac{p}{2}[\delta_x f_m + \delta_y g_m]$$

(5.3.1)

$$u_{m+1} = u_m - p[\delta_x f^*_{m+1} + \delta_y g^*_{m+1}]$$

where $\delta_x, \delta_y, \mu_x, \mu_y$ are defined by

$$\delta_x u^m_{ij} = u^m_{i+\frac{1}{2}j} - u^m_{i-\frac{1}{2}j}, \qquad \delta_y u^m_{ij} = u^m_{ij+\frac{1}{2}} - u^m_{ij-\frac{1}{2}}$$

$$\mu_x u^m_{ij} = \tfrac{1}{2}(u^m_{i+\frac{1}{2}j} + u^m_{i-\frac{1}{2}j}) \qquad \mu_y u^m_{ij} = \tfrac{1}{2}(u^m_{ij+\frac{1}{2}} + u^m_{ij-\frac{1}{2}})$$

However such modifications usually require the evaluation of the Jacobian matrices A, B, and therefore the simple form of the two-step formulation is destroyed. We mention the work of Lax and Wendroff [45] and Burstein [5], [6] in this context. In [62] Strang showed how the best stability characteristics $p|\lambda_A|, p|\lambda_B| \leq 1$ for an explicit scheme could be obtained. If we denote by $M_x$ the Lax Wendroff difference operator in the x-direction, then Strang's scheme consists of using

$$u_{m+1} = \tfrac{1}{2}[M_x M_y + M_y M_x]\, u_m$$

As pointed out by Burstein [5], this amounts to adding to the Lax-Wendroff operator in two dimensions the pseudo-viscous term

$$-\frac{1}{8}p^3\left\{(AB^2 + B^2A)(\Delta_x + \nabla_x)\Delta_y\nabla_y + (BA^2 + A^2B)(\Delta_y + \nabla_y)\Delta_x\nabla_x\right\} u_m$$

$$+ \frac{1}{8}p^4\left\{A^2B^2 + B^2A^2\right\} \Delta_x\nabla_x\Delta_y\nabla_y\, u_m$$

where $\Delta, \nabla$ are the usual forward and backward difference operators and where we have assumed for ease that A, B are constant matrices. We shall now derive a <u>multistep</u> version of Strang's scheme which requires nothing more than the evaluation of the functions f and g. Although this scheme requires more computing time than the normal Lax-Wendroff method (5.3.1), its improved stability characteristics will more than offset this disadvantage.

Consider, therefore, the two dimensional system of conservation laws

(5.3.2) $$\frac{\partial u}{\partial t} + \frac{\partial f}{\partial x} + \frac{\partial g}{\partial y} = 0$$

solved by the scheme

(5.3.3) $$v^{m+1}_{(1)} = \mu_y u_m - \frac{p}{2}\delta_y g_m$$

$$v^{m+1}_{(2)} = u_m - p\,\delta_y\, g^{m+1}_{(1)}$$

$$v^{m+1}_{(3)} = \mu_x v^{m+1}_{(2)} - \frac{p}{2}\delta_x f^{m+1}_{(2)}$$

$$v^{m+1}_{(4)} = v^{(m+1)}_{(2)} - p\,\delta_x\, f^{m+1}_{(3)}$$

(5.3.4) $$w^{m+1}_{(1)} = \mu_x u_m - \frac{p}{2}\delta_x f^m$$

$$w^{m+1}_{(2)} = u_m - p\,\delta_x\, f^{m+1}_{(1)}$$

$$w^{m+1}_{(3)} = \mu_y w^{m+1}_{(2)} - \frac{p}{2}\delta_y\, g^{m+1}_{(2)}$$

$$w^{m+1}_{(4)} = w^{m+1}_{(2)} - p\,\delta_y\, g^{m+1}_{(3)}$$

(5.3.5) $$u_{m+1} = \tfrac{1}{2}(v^{m+1}_{(4)} + w^{m+1}_{(4)})$$

where we have used the notation $g^{m+1}_{(1)} = g(v^{m+1}_{(1)})$ etc.

Since the majority of the computing time in such problems is taken up by the evaluation of the vector functions f and g, we will compare the algorithms on this basis. It is then easily seen that the multistep algorithm requires roughly twice as much computing time as does the Lax-Wendroff method (5.3.1), but the stability characteristics of (5.3.3, 5.3.4, 5.3.5) are $2\sqrt{2}$ times better than those of (5.3.1). Apart from this advantage the multistep Strang method is very easy to programme, requiring only a single procedure equivalent to a one dimensional Lax-Wendroff algorithm. Moreover, if it is required to add in a pseudo-viscous term, this merely takes its normal place in the respective one dimensional algorithm and will itself be a one-dimensional quantity. It would appear therefore that the multistep formulation of Strang's scheme has a great deal to offer in its simplicity and its stability.

The above results may be extended in a natural way to the more general systems of equations considered in Chapter 4. The two step Lax-Wendroff method for the system

$$\frac{\partial \underline{u}}{\partial t} + \frac{\partial \underline{f}(x, t, u)}{\partial x} = g(x, t, u)$$

may be written in the form

$$u^{m+1}_{(1)} = \mu_x u_m - \frac{p}{2}\left[\, \delta_x \bar{f}_{m+1} - h\, \bar{g}_{m+1} \right]$$

$$u_{m+1} = u_m - p\left[\, \delta_x \bar{f}^{m+1}_{(1)} - h\, \bar{g}^{m+1}_{(1)} \right]$$

where

$$\bar{f}_{m+1} = f(x, (m + \tfrac{1}{2})k, u_m)$$

and

$$\bar{f}^{m+1}_{(1)} = f(x, (m + \tfrac{1}{2})k, u^{m+1}_{(1)}) \quad \text{etc.}$$

Thus for the two dimensional system,

$$(5.3.6) \qquad \frac{\partial u}{\partial t} + \frac{\partial f}{\partial x}(x, y, t, u) + \frac{\partial g}{\partial y}(x, y, t, u) = \gamma(x, y, t, u),$$

we have the multistep Strang scheme

$$v^{m+1}_{(1)} = \mu_y u_m - \frac{p}{2}[\delta_y \bar{g}_{m+1} - h\bar{q}_{m+1}]$$

$$v^{m+1}_{(2)} = u_m - p[\delta_y \bar{g}^{m+1}_{(1)} - h\bar{q}^{m+1}_{(1)}]$$

$$v^{m+1}_{(3)} = \mu_x v^{m+1}_{(2)} - \frac{p}{2}[\delta_x \bar{f}^{m+1}_{(2)} - h\bar{s}^{m+1}_{(2)}]$$

$$v^{m+1}_{(4)} = v^{m+1}_{(2)} - p[\delta_x \bar{f}^{m+1}_{(3)} - h\bar{s}^{m+1}_{(3)}]$$

$$w^{m+1}_{(1)} = \mu_x u_m - \frac{p}{2}[\delta_x \bar{f}_{m+1} - h\bar{s}_{m+1}]$$

$$w^{m+1}_{(2)} = u_m - p[\delta_x \bar{f}^{m+1}_{(1)} - h\bar{s}^{m+1}_{(1)}]$$

$$w^{m+1}_{(3)} = \mu_y w^{m+1}_{(2)} - \frac{p}{2}[\delta_y \bar{g}^{m+1}_{(2)} - h\bar{q}^{m+1}_{(2)}]$$

$$w^{m+1}_{(4)} = w^{m+1}_{(2)} - p[\delta_y \bar{g}^{m+1}_{(3)} - h\bar{q}^{m+1}_{(3)}]$$

$$u_{m+1} = \tfrac{1}{2}v^{m+1}_{(4)} + w^{m+1}_{(4)})$$

where

$$(5.3.7) \qquad \zeta(x, y, t, u) = q(x, y, t, u) + s(x, y, t, u)$$

It is somewhat surprising that an arbitrary splitting of $\zeta$ in the form (5.3.7) maintains second order accuracy of the above scheme. Notice finally that we may allow $\zeta$ to include second order derivatives $u_{xx}$, $u_{xy}$, $u_{yy}$ and thus we can write the full Navier-Stokes equations in the form (5.3.6) (See Thommen [66]).

In this thesis finite difference methods have been presented for first order quasilinear hyperbolic systems of partial differential equations in one and two space dimensions which, when used with the appropriate technique for inserting the extra boundary data, have proved to be accurate and computationally efficient. It is hoped in the near future to use these methods for the solution of shocked problems in hydrodynamics in order to investigate the ability of the methods to represent shocks in physical problems.

## References

[1] Ames, W.F.: Non-linear Partial Differential Equations in Engineering. Mathematics in Science and Engineering, 18. Academic Press (1965).

[2] Arakawa, A.: Computational Design for Long term Numerical Integration of the Equations of fluid motion. Two dimensional Incompressible Flow, Part I. J. Comp. Phys. 1, 119-143, (1966).

[3] Bernstein, D.L.: Existence Theorems in Partial Differential Equations. Princeton University Press, (1950).

[4] Bulirsch and J. Stoer: Numerical Treatment of ordinary differential equations by extrapolation methods. Num. Math. 8, 1.-13, (1966).

[5] Burstein, S.Z.: Numerical calculations of multi dimensional shocked flows. A.I.A.A.J. 2, 211-217, (1964).

[6] Burstein, S.Z.: Finite difference calculations for hydrodynamic flows containing discontinuities. J. Comp. Phys. 1, 198-222, (1966).

[7] Carlson, B.G.: Solution of the Transport Equation by $S_n$ Approximations. Los Alamos Scientific Lab. Rept. LA - 1891, (1955).

[8] Conway, E. and J. Smoller: Global solutions of the Cauchy Problem for Quasilinear First Order Equations in several space variables. Comm. Pure Appl. Maths. 19, 95-105, (1966).

[9] Courant R. and K.O. Friedrichs: Supersonic Flow and Shock Waves. Interscience, New York, (1948).

[10] Courant, R. and K.O. Friedrichs and H. Lewy: On the partial differential equations of mathematical physics. New York Univ. Courant Inst. Math. Sci. Res. Rept. trans. Phyllis Fox, NYO-7687, (1956).

[11] Crocco, L.: A suggestion for the Numerical Solution of the Steady Navier-Stokes Equations. A.I.A.A.J. 3, 1824-1832, (1965).

[12] D'Jakonov, G. Ye: Difference schemes with a disintegrating operator for multidimensional problems. U.S.S.R. Comp. Math. and Math. Phys. 4, 581-607, (1963).

[13] Douglis, A.: The Continuous Dependence of Generalised Solutions of Non-linear Partial Differential Equations upon Initial Data. Comm. Pure Appl. Math. 14, 267-284, (1961).

[14] Friedrichs, K.O.: Symmetric Positive Linear Differential Equations Comm. Pure Appl. Math. 11, 333-418, (1956).

[15] Friedrichs, K.O.: Symmetric Hyperbolic Linear Differential Equations. Comm. Pure Appl. Math. 7, 345-392, (1964).

[16] Fromm, J.E.: Lagrangian difference approximations for fluid dynamics. Los Alamos Sci. Lab. Rept. LA-2535, (1961).

[17] Garabedian, P.R.: Partial Differential Equations. John Wiley and Sons, New York, (1964).

[18] Gary, J.: On certain finite difference schemes for hyperbolic systems. Maths. of Comp. 18, 1-18, (1964).

[19] Gourlay, A.R.: Alternating Direction Methods for Hyperbolic systems. Doctoral Thesis, University of St. Andrews, (1966).

[20] Gourlay, A.R. and A.R. Mitchell: Alternating Direction Implicit methods for hyperbolic systems. Num. Math. 8, 137-149, (1966).

[21] Gourlay, A.R. and A.R. Mitchell: Two level difference schemes for hyperbolic systems. S.I.A.M. Journal of Num. Anal. 3, 474-485, (1966).

[22] Gourlay, A.R. and A.R. Mitchell: A stable implicit method for hyperbolic systems in two space variables. Num. Math. 8, 367-375, (1966).

[23] Gourlay, A.R. and A.R. Mitchell: An Explicit Alternating direction method for hyperbolic systems.

[24] Gourlay, A.R. and J. Ll. Morris: Finite Difference Methods for Non-Linear Hyperbolic systems I, Maths. of Comp. (to appear

[25] Gourlay, A.R. and J. Ll. Morris: Finite Difference Methods for Non-Linear Hyperbolic Systems II, Maths. of Comp. (to appear).

[26] Gourlay, A.R. and J. Ll. Morris: A multistep formulation of the optimized Lax-Wendroff method for non-linear hyperbolic systems in two space variables. Maths. of Comp. (to appe

[27] Gragg, W.B.: Repeated Extrapolation to the limit in the Numerical Solution of Ordinary Differential Equations. Doctoral dissertation, University of California, Los Angeles, (1964).

[28] Gragg, W.B.: On Extrapolation Algorithms for Ordinary Initial Value Problems. J.S.I.A.M. Numer. Anal. Ser. B, 2, No. 3, 384-403, (1965).

[29] Gragg, W.B. and H.J. Stetter: Generalized Multistep predictor-corrector method J. Assoc. Comput. Math. 11, 188-209, (1964).

[30] Hedstrom, G.W.: The near-stability of the Lax-Wendroff method. Num. Math. 7, 73-77, (1965).

[31] Heie, H. and D.C. Leigh: Numerical Stability of Hyperbolic Equations in three Independent Variables. A.I.A.A.J. 3, 1099-1103, (1965).

[32] Hofmann, P.: Asymptotic Expansions of the Discretezation Erro of Boundary Value Problems of the Laplace Equation in Rectangular Domains. Num. Math. 9, 302-322, (1967).

[33] Jeffrey, A.: The Development of Jump Discontinuities in Non-linear Hyperbolic Systems of Equations in Two Independent Variables. Arch. Rat. Mech. Axal. 14, 27-37, (1963).

[34] Jeffrey, A.: The Breaking of Waves on a Sloping Beach. ZAMP, 15, 97-106, (1964).

[35] Jeffrey, A.: On a Class of Non-Breaking Finite Amplitude Water Wave. ZAMP, 18, 57-65, (1967).

[36] Jeffrey, A. and T. Tanuiti: Non Linear Wave Propagation. Mathematics in Science and Engineering, 9, Academic Press, (1964).

[37] Keller, H.B. and B. Wendroff: On the Formulation and Analysis of Numerical Methods for time Dependent Transport Equations. Comm. Pure Appl. Math. 10, 567-582, (1957).

[38] Kreiss, H.O.: Über die stabilitats depösition fur differenzgleichungen die partielle differenzgleighungen approximieren. BIT, 153-181, (1962).

[39] Kreiss, H.O.: Über implizite Differenzmethoden fur partielle Differenzgleichungen. Num. Math. 5, 24-47 (1963).

[40] Kreiss, H.O.: On difference approximations of dissipative type for hyperbolic differential equations. Comm. Pure. Appl. Math. 17, 335-353 (1964)

[41] Lax, P.D.: Non-linear Hyperbolic equations. Comm. Pure Appl. Math. 6, 231-258, (1953).

[42] Lax, P.D.: Hyperbolic systems of Conservation Laws, II. Comm. Pure Appl. Math. 10, 537-566, (1957).

[43] Lax, P.D.: On the stability of Difference Approximations to Solutions of Hyperbolic Equations with Variable Coefficients. Comm. Pure Appl. Math. 14, 497-520, (1961).

[44] Lax, P.D. and R.D. Richtmyer: Survey of the Stability of Linear Finite Difference Equations. Comm. Pure Appl. Maths. 9, 267-293, (1956).

[45] Lax, P.D. and B. Wendroff: Systems of Conservation Laws. Comm. Pure Appl. Math. 13, 217-237, (1960).

[46] Longley, H.J.: Methods of differencing in Eulerian hydrodynamic Los Alamos Sci. Lab. Rept., LA-2379, (1960).

[47] Lundquist, S.: Studies in magnetohydrodynamics. Arkiv Fysik, 5, 297-301, (1952).

[48] Marchuk, G.I.: Numerical Methods for solving Weather Forecastin and Climate Theory Problems - Part I. Research Translation Air Force Cambridge Res. Labs. TR.520.

[49] Mayers, D.F.: The deferred approach to the limit in ordinary differential equations. Comput. Journal, 7, 54-57, (1964).

[50] W.F. Noh and M.H. Protter: Difference Methods and the Equations of Hydrodynamics. J. of Math. and Mech. 12, 149-192, (1963).

[51] O'Brien, Hyman and Kaplan: A study of the numerical solution of partial differential equations. J. Maths. Phys. 29, 223-251, (1951).

[52] Parter, S.V. Stability, Convergence, and pseudo-stability of finite difference equations for an over determined problem. Num. Math. 4, 277-292, (1962).

[53] Peaceman , D.W. and H.H. Rachford: The numerical solution of parabolic and elliptic differential equations. J.S.I.A.M., 3, 28-41, (1955).

[54] Pike, H.M.: Hyperbolic Partial Differential Equations. Numerical Analysis. An Introduction. Ed. J. Walsh, Academic Press (1966).

[55] Ralston, A.: A First Course in Numerical Analysis. McGraw-Hill, (1965).

[56] Richardson, L.F. and J.A. Gaunt: The deferred approach to the limit II - Interpenetrating lattices. Philo. Trans. Roy. Soc. London, Ser. A., 226, 350-361, (1927).

[57] Richtmyer, R.D.: Difference Methods for Initial Value Problems. Interscience.

[58] Richtmyer, R.D.: A survey of Difference methods for Non-Steady Fluid Dynamics. NCAR Technical Notes 63-2.

[59] Richtmyer, R.D. and K.W. Morton: Stability Studies for difference equations. New York University Courant Inst. Math. Sci. Res. Rept. NYO - 1480 - 5 (1964).

[60] Sauerwein, H.: Numerical Calculations of Multidimensional and Unsteady Flows by the method of Characteristics. J. Comp. Phys. 1, 406-432, (1967).

[61] Strang, G.: Accurate Difference Methods I. Linear Cauchy Problems. Arch. Rat. Mech. Anal. 12, 392-402, (1963).

[62] Strang, G.: Accurate Difference Methods II. Non-linear Problems. Num. Math. 6, 37-46, (1964).

[63] Stetter, H.J.: Asymptotic expansion for the error of discretization. Algorithms for Non-linear Functional Equations. Num. Math. 7, 18-31, (1965).

[64] Thomée, V.: Stability of Difference Schemes in Maximum-Norm. Journal of Differential Equations, 1, 273-292, (1965).

[65] Thomée, V.: A stable difference scheme for the mixed boundary problem for a hyperbolic first-order system in two dimensions. J.S.I.A.M. 10, 229-245, (1962).

[66] Thommen, H.U.: Numerical Integration of the Navier-Stokes Equations. Z.A.M.P. 17, 369-384, (1966).

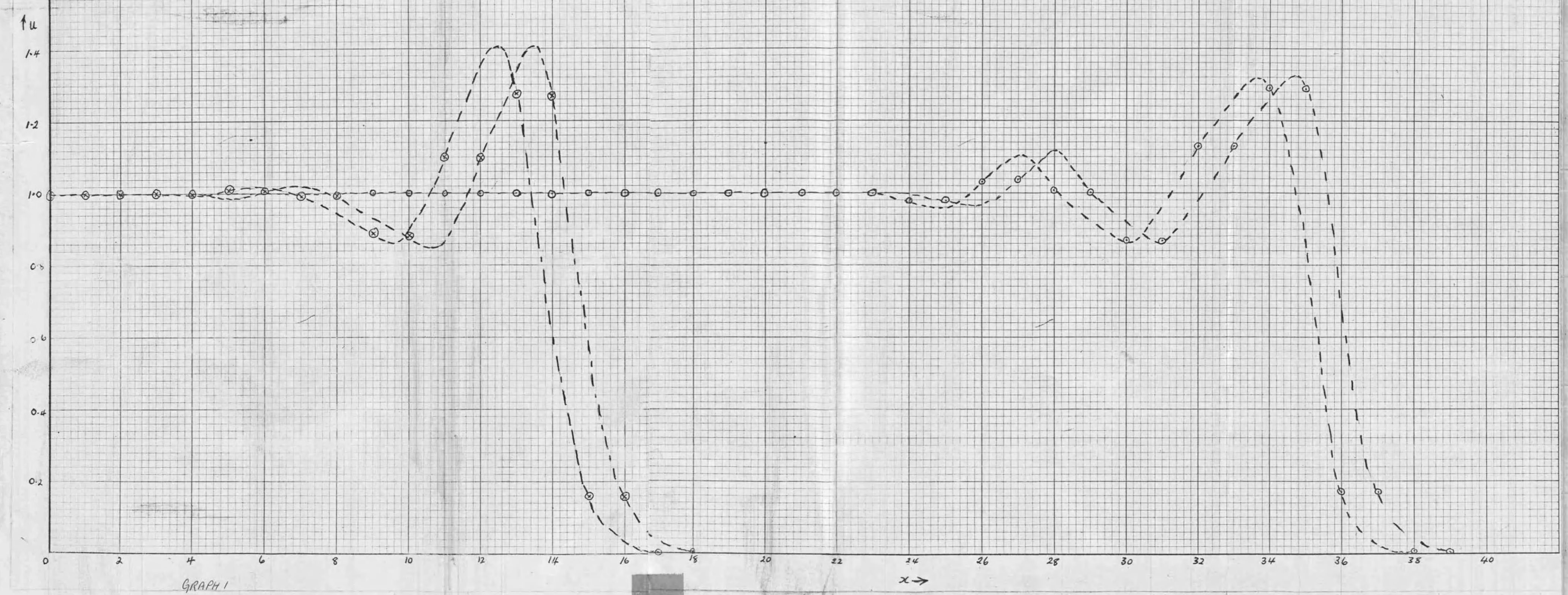
[67] Thommen, H.U.: A method for the Numerical Solution of the Complete Navier-Stokes Equations for Steady Flows. General Dynamics / Convair Rept. GDC - ERR - AN733.

[68] Varga, R.S.: Matrix Iterative Analysis. Prentice Hall, (1962).

[69] Von Neumann, J. and R.D. Richtmyer: A method for the numerical calculations of hydrodynamic shocks. J. Appl. Phys. 21, 232-237, (1950).

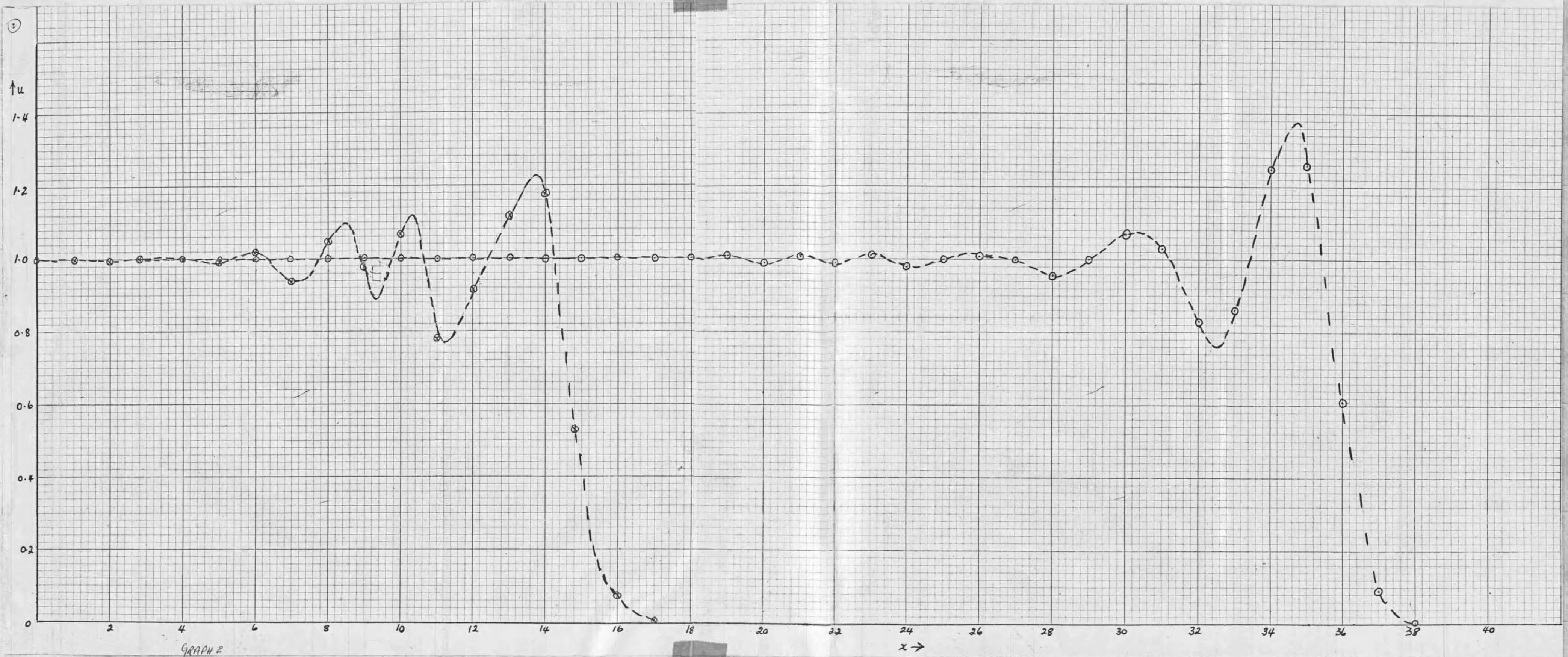
[70] Gourlay, A.R. and J. Ll. Morris: Deferred approach to the limit in non-linear hyperbolic systems. Computer Journal (to appear).

GRAPH 1


u
1·4
1·2
1·0
0·8
0·6
0·4
0·2
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
x →

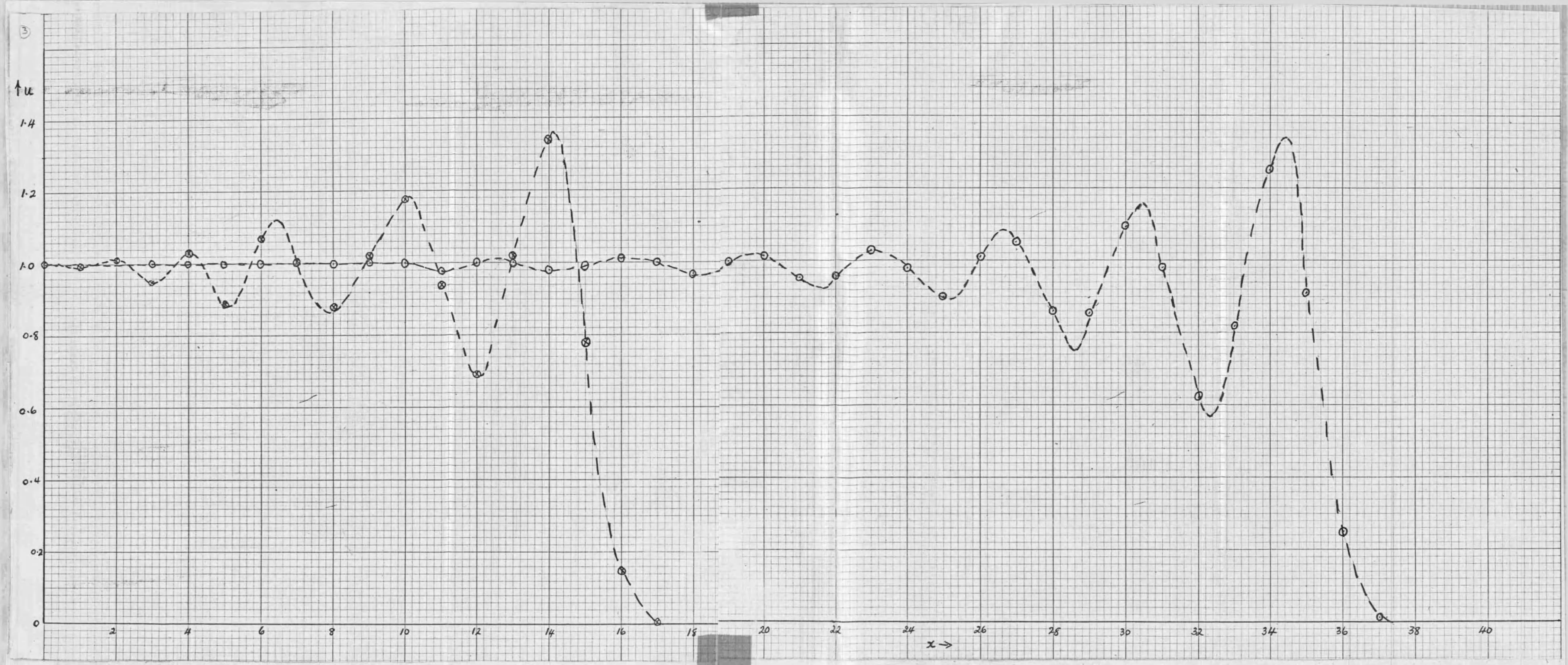
GRAPH 1

GRAPH 2


②
↑u
1·4
1·2
1·0
0·8
0·6
0·4
0·2
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
x →

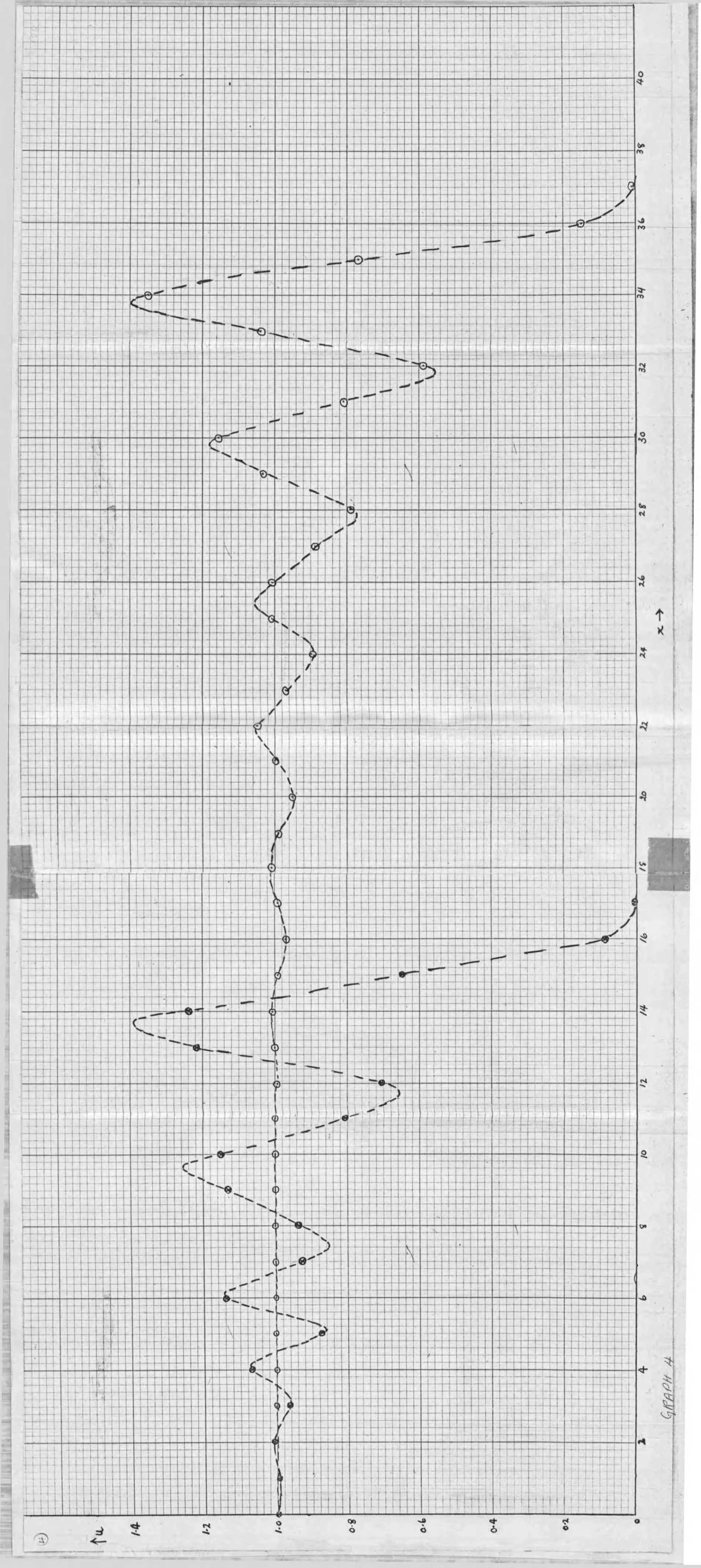
GRAPH 2

GRAPH 3


③
↑u
1·4
1·2
1·0
0·8
0·6
0·4
0·2
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
x →

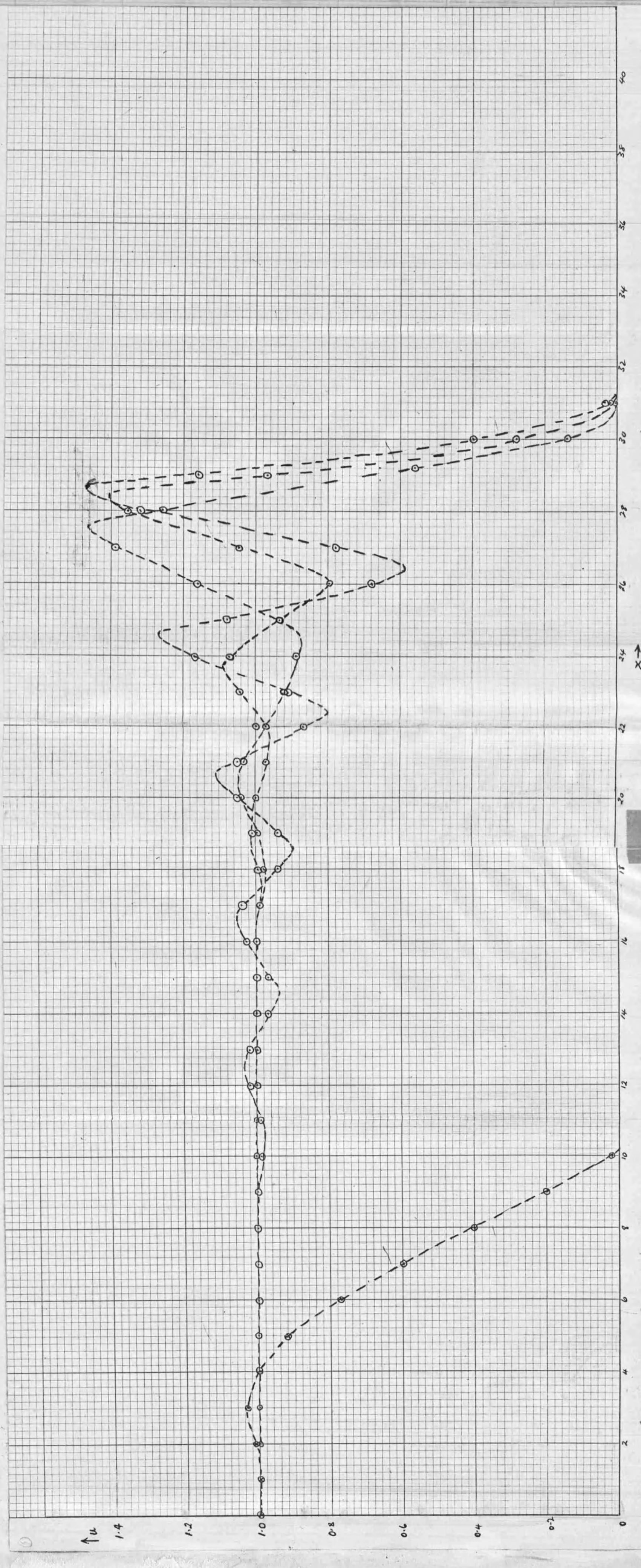
GRAPH 3

GRAPH 4


↑u
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
x →

GRAPH 4

GRAPH 5.


u
1·4
1·2
1·0
0·8
0·6
0·4
0·2
0
x
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40

GRAPH 5.